DOI: 10.53469/jssh.2025.7(12).13

# Financial English Lexical Threshold Range-aided Analysis on Citibank & Standard Chartered Bank's Annual Reports 2011-2020

**Dan Zhang[1],*, Xiaowen Gu[2], Xingye Pan[3]**

[123]College of Foreign Languages, Shanghai Maritime University, Pudong, Shanghai, China
*[1]danzhang@shmtu.edu.cn, [2]shhaishi66@163.com, [3]202210810107@stu.shmtu.edu.cn*
**Correspondence Author*

**Abstract:** ***Financial English, a key branch of English for Specific Purposes (ESP), has become increasingly important in the context of economic globalization. Its lexical threshold in China has not yet been quantified, but a corpus-driven approach makes it possible. After discussing the relationships among sight vocabulary, lexical coverage, word frequency, lexical threshold, and reading comprehension, this paper preprocesses Citibank and Standard Chartered Bank annual reports from 2011–2020 and analyzes 1,152,132 tokens using Range. The research indicates that to reach the minimum lexical coverage of 95%, learners should master approximately 4,000 word families. To achieve a more desirable lexical coverage of 98%, learners should not only acquire 4,565 word families from the BNC corpus but also an additional 274 key Financial English words beyond the BNC 15-level word list. The author suggests using these two datasets as benchmarks for the lexical threshold of Financial English learners, aiming to provide a lexical-level basis for vocabulary teaching, syllabus design, wordlist compilation, and textbook selection.***

**Keywords:** Lexical threshold, Lexical coverage, Word frequency, Financial English.

## 1. Introduction

Lexicology research has evolved from describing individual words to analyzing the overall structure of the vocabulary system with a data-driven method. However, the concept of professional vocabulary remains insufficiently defined, which makes it difficult to measure its accurate density in texts, to set lexical thresholds, and to design targeted teaching strategies based on its unique characteristics. Corpus-based lexical analysis plays a pivotal role in accurately identifying professional vocabulary and refining both its theoretical description and pedagogical application.

The pioneering work of Chung and Nation (2003) established a systematic method for identifying professional vocabulary, guiding subsequent studies on business and financial English. Vocabulary size, lexical coverage, and reading comprehension are closely linked, and lexical coverage has become the primary criterion for determining lexical thresholds. Although comprehension is also influenced by other reasons, coverage remains the most influential factor. Based on the experimental corpus of general English, Nation (2001), Laufer (2010) and Schmitt (2011) drew similar conclusions: 95% lexical coverage requires mastering roughly 4,000 – 5,000 word families, and 98% coverage requires mastering about 8,000–9,000 word families in general English. The vocabulary corresponding to 95% and 98% lexical coverage is in the range of 2,500-5000 and 5,000 - 9,000 word families, respectively.

Nation (2006) further suggests that many disciplines contain 1,000–2,000 domain-specific word families. Whether the financial English threshold can be defined by merely adding 1,000-2,000 words to the general English is still unclear, and no study has yet measured the lexical threshold of financial English using bank annual reports. Given that financial English has been included in the Ministry of Education's undergraduate curriculum, establishing its lexical threshold is essential for advancing both theoretical research and instructional practice.

The purpose of this study is to identify the vocabulary size required to achieve 95% and 98% lexical coverage in financial English texts, thereby providing an empirical reference for setting the lexical threshold of Financial English. By building and analyzing a corpus of Citibank and Standard Chartered Bank annual reports from 2011 to 2020 with the help of Range software, this study analyzes word frequency distribution and coverage to determine the vocabulary threshold of Financial English.

The findings can also offer practical value for curriculum design, vocabulary instruction, syllabus development, wordlist compilation, and textbook selection, contributing to a clearer understanding of the vocabulary demands faced by learners of financial English.

Based on the relationship among vocabulary size, lexical coverage, and reading comprehension, this study aims to answer the following questions:

1) How many vocabularies should learners master in order to reach the minimum lexical coverage and the relatively abundant lexical coverage?

2) Whether the lexical threshold for financial English is the same as general English. If not, what are the reasons?

3) Whether the lexical threshold of Financial English can be determined simply by adding 1,000 – 2,000 word families to that of General English, as suggested by Nation (2006). If not, what are the reasons?

4) Whether the word frequency of financial English is the same as general English. If not, what are the reasons?

## 2. Literature Review

Laufer (1989) attempted to determine the minimum lexical coverage for comprehension of authentic texts. She found that the students who could understand 95% of the text words scored an acceptable score for her experiment. In a further study, she tried to ascertain what vocabulary size is required by learners to achieve an acceptable score in a reading comprehension test (Laufer, 1992). The results showed that a vocabulary size of 3,000 word families – or 4,800 lexical items – predicts a minimum reading comprehension test result of an acceptable score. The two studies therefore suggest that a lexical coverage of 95% in a text, or knowledge of a vocabulary of 3,000 word families, is the much desirable threshold for learners to be able to achieve satisfactory reading comprehension. However, Hirsh and Nation (1992) suggested that 98% lexical coverage is the desirable level for enjoyable reading. Hu and Nation (2000), testing a fiction text at four coverage levels, found that while some learners achieved adequate comprehension at 95% coverage, most did not. They concluded that learners need to know 98% of the running words to read for pleasure, supporting the 98% coverage recommendation of Hirsh and Nation.

Laufer (2010) further proposed two lexical thresholds: the lowest lexical threshold and the threshold of relative abundance. The ideal lexical threshold is to master 8,000 - 9,000 word families to achieve 98% lexical coverage, while the lowest lexical threshold is to master 4,000-5000 word families to achieve 95% lexical coverage. These findings are well-supported by lexicological research.

Nation (2006) takes novels and newspapers as the corpus, and the result is mastering 4,000 word families to reach 95% lexical coverage and 8,000-9,000 word families to reach 98% lexical coverage. Hsu (2011) takes business textbooks and business research papers as the corpus. The finding is mastering 3,500 - 5,000 word families to reach 95%-98% lexical coverage and 5,000-8,000 word families to reach 95%-98% lexical coverage separately.

Students' reading comprehension abilities are largely affected by their lexical level. Laufer and Sim (1985), for example, while investigating how students with partial linguistic knowledge interpret a text, concluded that a strong vocabulary base is essential for learner to use contextual clues effectively. In another study, Laufer and Sim (1985) found that learners first relied on word meaning, then on subject knowledge, and finally on syntax when attempting to comprehend texts. In their research, Ulijn and Strother (1990) suggest that reading depends chiefly on understanding the meaning of the text's words and its subject matter. Laufer (1992) conducted a study in order to find out whether the learners' general academic ability or their lexical level is a better predictor of reading, and her results indicated that it is the learner's lexical level which could help predict their reading ability. Finally, Haynes and Baker (1993), in an experiment with college students using a text containing both linguistic and non-linguistic lexical elements, concluded that readers require not only effective reading strategies but also extensive vocabulary knowledge for comprehension. Thus, vocabulary is a fundamental factor in understanding a text. Chinese students preparing for university studies in Financial English are no exception. They must acquire sufficient vocabulary to engage effectively in their courses; however, the precise amount required for successful comprehension remains uncertain.

Hsu (2011) summarizes the research data of lexicology: the high-frequency 1000 level vocabulary accounts for about 78% - 81% of the vocabulary coverage of general texts and the high-frequency 2000 level vocabulary accounts for 8% - 9%; 3000 words account for another 3% - 5%. The 4,000 - 5,000 word level of intermediate frequency accounts for another 3% and the 6,000 - 9,000 word level of intermediate frequency accounts for another 2%. The 10,000 - 14,000 word level of low frequency only accounts for 1% of the text. Mature vocabulary analysis software can automatically analyze and process the vocabulary status of reading text and provide detailed data support for researchers.

## 3. Research Methodology

### 3.1 Data Collection

The corpus in this paper is the annual reports of Citibank & Standard Chartered Bank from 2011 to 2020. Citibank is one of the largest banks in the United States and a leading foreign bank in China. While Standard Chartered Bank holds a leading position among international banks operating in the country. Therefore, using the annual reports of these two banks as the corpus provides a valuable reference for setting the lexical threshold.

The reports were downloaded from the official websites of these two companies, with a total of 20 articles and 1,244,458 tokens. The selected corpus has a large time span and covers the financial information of each period.

Citibank's Annual reports include an overview of the past year's information, management discussions, financial statements, balance sheet review, risk reporting, and corporate information, offering representative and standardized financial texts.

While the annual reports of Standard Chartered Bank are mainly separated into five parts: group overview, operating and financial review, corporate governance, financial statements and notes, and supplementary information.

### 3.2 Research Tools

The analysis tool used in this experiment is the Range program (BNC version) developed by

Paul Nation (free download website: http://www.victoria.ac.nz/lals/about/staff/paul-nation). This version extracts 14,000 word families from the British National Corpus (BNC), which contains 100 million words, and classifies them into 14 frequency levels, with each level representing 1,000 word families. These levels were developed based on the principle that learners typically acquire vocabulary in order of descending frequency. In addition, the BNC includes a proper noun list 15 and an exclamation list 16. The corpus contains over 100 million words from 4,124 modern British English texts, of which 90% are written texts and 10% are spoken language materials.

### 3.3 Research Procedures

In this part, the author introduces the procedures of data preprocessing and how the data were processed in the study.

3.3.1 Data Preparation

Before using Range for statistical analysis of the above corpus, the author preprocesses the self-built corpus to accurately obtain relevant data and ensure that Range analysis eliminates possible interference factors in the corpus. The preprocessing mainly includes the following points:

1) Using the spelling check function in Microsoft Word, all American English words in the corpus were converted into British English, ensuring consistency with the vocabulary of the British National Corpus. Otherwise, Range might incorrectly classify some American English words — originally belonging to the top 14 frequency levels—into the "not in the lists" category.

2) The corpus contains numerous hyphenated compound words (e.g., retailer-friendly, post-tax). Although most components of these compounds fall within the top 14 frequency levels, the Range software categorizes such forms as "symbols out of the list." Since learners can generally infer their meanings, hyphens were replaced with spaces in batches before the Range software analyzes the corpus.

3) Compound words with transparent meanings and no hyphens, such as midsize and voicemail, were removed from the corpus. If the components of these compounds appear in level 14, they should not be treated as low-frequency words beyond level 14.

The total number of shape characters after preprocessing is 1,152,132.

3.3.2 Data Input and Processing

Store the preprocessed text in the folder where the Range BNC version is located for analysis and statistics. The steps are as follows:

1) Run the Range32.exe program and set the relevant parameters of the program. Change the number of "Number of Baseword Files" from the preset 3 to 15, that is, use the first 15 of the 16 baseword files built in the software for analysis (the 16th baseword file is an exclamation word family).

2) Open the processed file through the menu "file" - "open", and then click "Process Files" at the bottom of the left window to complete the operation.

## 4. Results And Discussion

### 4.1 Results

The analysis file generated by range is stored in the Range program folder, and the end of the file name is_ Range, the contents contained in the document are the analysis results of 20 annual reports of Citibank and Standard Chartered Bank from 2011 to 2020 (see Table 1).

**Table 1:** Analysis on results of Range BNC

| WORD LIST | TOKENS/% | TYPES/% | FAMILIES |
|---|---|---|---|
| One | 807195/70.06 | 2913/17.91 | 916 |
| Two | 151968/13.19 | 1908/11.73 | 716 |
| Three | 31993/ 2.78 | 884/ 5.37 | 451 |
| Four | 48104/ 4.18 | 858/ 5.29 | 430 |
| Five | 22429/ 1.95 | 590/ 3.64 | 321 |
| Six | 12305/ 1.07 | 395/ 2.40 | 240 |
| Seven | 9347/ 0.81 | 292/ 1.77 | 199 |
| Eight | 3648/ 0.32 | 220/ 1.35 | 166 |
| Nine | 3004/ 0.26 | 173/ 1.06 | 122 |
| Ten | 1223/ 0.11 | 128/ 0.79 | 104 |
| 11 | 4027/ 0.35 | 98/ 0.62 | 77 |
| 12 | 973/ 0.08 | 81/ 0.51 | 70 |
| 13 | 871/ 0.08 | 72/ 0.45 | 61 |
| 14 | 921/ 0.08 | 62/ 0.39 | 51 |
| 15 | 14684/ 1.27 | 641/ 3.87 | 641 |
| not in the lists | 39440/ 3.42 | 3587/42.86 | ????? |
| Total | 1152132 | 12902 | 4565 |

### 4.2 Discussion

Range analyzed 1,152,132 tokens in the sample texts of Citibank & Standard Chartered Bank's annual reports from 2011 to 2020. The experimental results show that 3,273 word families (916 families of word list 1+716 families of word list 2+451 families of word list 3+430 families of word list 4+321 families of word list 5+240 families of word list 6+199 families of word list 7) plus 641 proper nouns in the fifteenth word list can achieve 95.31% lexical coverage, namely 3914 families: 70.06% of word list 1 + 13.19% of word list 2 + 2.78% of word list 3 + 4.18% of word list 4 +1.95% of word list 5 +1.07% of word list 6 +0.81% of word list 7 + 1.27% of word list 15 equals 95.31%. However, even 4565 word families can not reach 98% lexical coverage: 70.06% of word list 1 + 13.19% of word list 2 + 2.78% of word list 3 + 4.18% of word list 4 + 1.95% of word list 5 + 1.07% of word list 6 + 0.81% of word list 7 + 0.32% of word list 8 + 0.26% of word list 9 + 0.11% of word list 10 + 0.35% of word list 11 + 0.08% of word list 12 + 0.08% of word list 13 + 0.08% of word list 14 + 1.27% of word list 15 equals 96.59%. Based on 95% and 98% lexical coverage, the minimum lexical threshold for financial English is 3,914 word families but the relatively abundant lexical threshold is over 4,565 word families.

4.2.1 Minimum Lexical Threshold

The minimum lexical threshold of financial English identified in this experiment closely aligns with that of general English, approximately 4,000 word families for 95% lexical coverage. Nation (2006) suggests that specialized fields typically require an additional 1,000 – 2,000 word families. This place the lexical threshold for professional English at 5,000–6,000 families. However, the present findings indicate that the threshold for financial English is nearly identical to that of general English, contradicting Nation's assumption.

In this regard, the author analyzes this as follows: financial English vocabulary is a part of business English vocabulary. According to Li (2015), business English vocabulary can be divided into sub-professional and professional types. "Sub-professional business English vocabulary" appears in both business and non-business contexts, carrying both specialized and general meanings. Formally, it is indistinguishable from general English vocabulary, but when used in business communication or specific commercial contexts, it acquires specialized meanings. For instance, draft means “manuscript” in general English but refers to “a bill issued by a bank for cashing a remittance” in business English. As sub-professional vocabulary constitutes the majority of business English vocabulary, the lexical level of business English—and by extension financial English—does not significantly differ from that of general English for achieving 95% lexical coverage. Therefore, it is inappropriate to simply add 1,000 – 2,000 word families to the general English threshold as the minimum lexical threshold for financial English.

### 4.2.2 Relatively Abundant Lexical Threshold

On the other hand, it is questionable that even 4,565 word families fail to achieve 98% lexical coverage. The key to the problem is the words that not in the lists. Upon careful examination, the author found that many high-frequency terms were incorrectly classified under this category. For instance, GAAP—the abbreviation of Generally Accepted Accounting Principles—is a core item of financial English vocabulary. Details are provided in Appendices 1, 2, and 3.

In total, 274 types and 16,266 tokens were identified as “not in the lists,” including 192 abbreviations, 12 well-known financial institutions, and 70 other finance-related terms. The author argues that these financial English terms, unrecognized by the BNC, constitute sight vocabulary that learners must master, as they significantly influence reading comprehension in financial English.

Firstly, the abbreviations include common monetary unit like AUD (Australian Dollar), CNY (Chinese Yuan), EUR (European Dollar), GBP (Great Britain Pound), HKD (Hong Kong Dollar), INR (Indian Rupees), JPY (Japanese Yen), PKR (Pakistani Rupee), etc. These monetary units are commonly used in foreign exchange transactions so it is necessary to master them for financial English learning.

Secondly, there are many financial organizations and banks in the abbreviations list. For example, ADB (Asian Development Bank), AEB (American Express Bank), ALCO (Asset and Liability Committee), ASE (American Stock Exchange), BCBS (Basel Committee on Banking Supervision), PWC (Price Waterhouse Coopers), etc.

In addition, there are many common financial English vocabularies. For example, AML (Anti-Money Laundering) refers to a comprehensive system in which governments, through legislative and judicial means, mobilize relevant organizations and financial institutions to identify, manage, and penalize money-laundering activities, thereby preventing criminal conduct. CFFO (Cash Flow from Operations) denotes cash flow generated from business activities, representing a company’s main source of funds and reflecting its actual operating performance.CG (Credit Grade) refers to the rating categories assigned by credit evaluation agencies based on the results of enterprise credit assessments. It reflects a company’s credit level. Western countries often use rating systems such as AAA, AA, and A, or the three-tier, nine-grade system (3A–3C) to guide investor decisions.

The vocabulary in Appendix Two consists of well-known financial institutions. For instance, the largest Europe’s insurance company and one of the world’s leading insurance and asset management groups—Allianz. Temasek Holdings, established in 1974, is the most prominent of several companies wholly owned by the Singapore government. It controls nearly all of Singapore’s major enterprises with the highest turnover.

The vocabulary in Appendix Three includes other common financial English terms. For instance, amalgamation refers to the process by which two or more enterprises merge to form a new entity by integrating their assets in accordance with relevant laws and regulations. Another example is repos, short for repurchase agreements, which involve selling a security with an agreement to repurchase it at a predetermined price on a future date. Repo transactions are a key tool used by central banks to regulate market liquidity and may occur between financial or non-financial institutions.

After calculation, the lexical threshold of these words reaches 1.41%, as there are 39,440 tokens classified as “not in the lists.” Adding this 1.41% to the 96.59% lexical threshold yields a total of 98%. Therefore, to achieve a relatively sufficient lexical coverage of 98%, learners need to master not only the 4,565 word families in the BNC corpus but also 274 key financial English terms beyond its 15 levels. The corpus used—annual reports of Citibank and Standard Chartered Bank from 2011 to 2020—thus meets learners’ minimum and relatively adequate lexical threshold.

As mentioned earlier, lexical thresholds vary by discipline and text type. As shown in Table 1, business textbooks require 5,000 word families to reach 98% coverage, which is most comparable to this corpus. The present study’s result—4,565 word families plus 274 types—is slightly lower than that of business texts. This may be because the corpus comprises annual reports from Citibank and Standard Chartered Bank, which include not only specialized financial vocabulary such as consolidated income statement, cash flow statement, and balance sheet, but also general English expressions such as chairman’s statement, shareholder information, and statement of directors’ responsibilities. Consequently, its lexical specialization is lower than that of business textbooks.

### 4.2.3 Word Frequency

According to Hsu (2011), the high-frequency vocabulary of general English accounts for 89%-95% lexical threshold, the intermediate-frequency vocabulary accounts for about 5% and the low-frequency vocabulary accounts for about 1%. However, according to Table 2, the high-frequency vocabulary of financial English accounts for 86.03%, the intermediate-frequency vocabulary accounts for 8.59% and the low-frequency vocabulary accounts for 0.7%.

The comparison shows that high-frequency vocabulary in financial English is lower than in general English, due to the corpus’s high degree of specialization and extensive use of technical terms.

In contrast, the intermediate-frequency vocabulary in financial English is considerably higher than in general English. Schmitt and Schmitt (2014) emphasized the crucial role of mastering intermediate-frequency vocabulary for fluent reading. They argued that achieving 95–98% lexical coverage requires not only high-frequency and specialized vocabulary but also a substantial amount of intermediate - frequency vocabulary. Therefore, intermediate-frequency words play a particularly important role in reading financial English texts.

The low-frequency vocabulary of financial English is comparable to that of general English and thus requires no further analysis. The focus here is on words not included in the lists. According to Appendices One, Two, and Three, IFRS (International Financial Reporting Standards) has the highest frequency, appearing 840 times, followed by amortised (532 times), CBS (Classified Balance Sheet, 345 times), and CFA (Cash Flow Accounting, 321 times). In total, 46 words occur more than 100 times. These high-frequency financial terms are essential for achieving effective reading comprehension.

## 5. Conclusion

Based on the 95%–98% lexical coverage and the annual reports of Citibank and Standard Chartered Bank over the past decade, this study shows that the minimum lexical threshold for financial English is 3,914 word families, while the relatively abundant threshold is about 4,565 word families and 274 types. The latter corresponds to a reading accuracy of 70%; if higher accuracy is required, the lexical threshold should increase accordingly (Nation, 2013). The minimum lexical threshold for financial English is not achieved by simply adding 1,000–2,000 technical terms to that of general English, since sub-professional vocabulary constitutes the majority of business English. The relatively abundant lexical threshold of financial English is slightly lower than that of business texts, mainly because many high-frequency words, such as GAAP, fall outside the BNC lists. Most importantly, the 274 types—comprising abbreviations, well-known financial institutions, and other key financial terms—require recognition of their meanings within financial contexts. High-frequency general English vocabulary is not applicable to financial English, while intermediate-frequency vocabulary is crucial for reading comprehension. Therefore, more emphasis should be placed on mastering intermediate-frequency financial English words.

The relatively abundant lexical threshold of financial English is much higher than that of general English, mainly due to its distinctive characteristics. Financial English contains numerous abbreviations classified as “not in the lists.” These sight vocabularies—particularly the 274 types outside the lists—pose significant obstacles to reading comprehension and therefore warrant focused attention. High-frequency vocabulary in financial English accounts for 86.03%, intermediate-frequency for 8.59%, and low-frequency for 0.7%. The proportion of high-frequency words is lower than in general English because of the corpus’s high level of specialization. In contrast, the share of intermediate - frequency vocabulary is considerably higher, indicating that achieving 95%–98% lexical coverage requires not only high-frequency and technical terms but also mastery of a substantial amount of intermediate-frequency vocabulary. Low-frequency vocabulary shows a pattern similar to that of general English. Most importantly, many high-frequency financial terms fall into the “not in the lists” category and must be emphasized to ensure smooth reading comprehension.

The corpus is composed of the 2011–2020 annual reports of Citibank and Standard Chartered Bank, with a primary focus on banking-related content. Given that the financial industry encompasses diverse sub-sectors such as insurance, leasing, and securities, corpora derived from these domains can also serve as valuable resources for future research of financial English. The analysis tool in this study is the Range program (BNC version), which focuses on lexical coverage and word frequency. However, it lacks functions for rapid semantic feature extraction and analysis of lexical distribution, density, and complexity. Therefore, future research could benefit from fine-grained retrieval and classification by part of speech, grammar, and semantics to compile a comprehensive and detailed financial English word list.

## References

[1] Hsu, W. (2011) The vocabulary thresholds of business textbooks and business research articles for EFL learners. *English for Specific Purpose*, 30 (4), 247-257.

[2] Hsu, W. (2014) Measuring the vocabulary load of engineering textbooks for EFL undergraduates. *English for Specific Purposes*, 33 (1), 54-65.

[3] Hu, M., & Nation, I. S. P. (2000) Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13 (1), 403-430.

[4] Laufer, B. (1992) How much lexis is necessary for reading comprehension. *Vocabulary and Applied Linguistics*, P. Arnaud, London: Macmillan.

[5] Laufer, B. (2010) Lexical threshold revisited: Lexical text coverage, learners’ vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22 (1), 15-30.

[6] Laufer, B. (2013) Lexical thresholds for reading comprehension: What they are and how they can be used for teaching purposes. *TESOL*, 47, 867-872.

[7] Nation, I. S. P. (2001) *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

[8] Nation, I. S. P. (2006) How large a vocabulary is needed for reading and listening. *The Canadian Modern Language Review*, 63 (1), 59-82.

[9] Nation, I. S. P. (2006) How large a vocabulary is needed for reading and listening. *The Canadian Modern Language Review*, 63, 59-82.

[10] Nikolaos,K. (2007). Creating a Business Word List for Teaching Business English. *ELIA*, 7, 79-102.

[11] Schmitt, N., Jiang, X., & Grabe, W. (2011) *The percentage of words known in a text and reading comprehension*. *Modern Language Journal*, 95 (1), 26-43.

[12] Schmitt, N., Jiang, X., & Grabe, W. (2011) The percentage of words known in a text and reading comprehension. Modern Language Journal, 95 (1), 26-43.

[13] Schmitt, N., & Schmitt, D. (2014) A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, 47 (4), 484-503.