The Exploration of Challenges and Applications of Corpus Linguistics in the Compilation of Health-Related Local Wisdom Vocabulary

Chatchanok Hengsuko
Udon Thani Rajabhat University, Thailand
E-mail: chatchanok.he@udru.ac.th

****

Abstract

This study explores the challenges and applications of corpus linguistics in compiling a specialized English vocabulary list for health-related local wisdom. The research highlights the increasing importance of preserving and disseminating local wisdom, particularly in health contexts, to foster global recognition and academic collaboration. However, a critical problem persists: the lack of a standardized English vocabulary resource for health-related local wisdom, which impedes international communication and limits the accessibility of Thai health knowledge. This qualitative study analyzed 30 articles from the Thai Journal Citation Index (TCI) database using tools such as WordSmith Tool and AntWordProfiler to examine word frequency and vocabulary profiles systematically.

The findings reveal that 60.24% of the identified words belong to the first 1,000 most common English words, 9.64% are academic English vocabulary, and 6.66% fall within the second 1,000 most common words, while 23.46% constitute other terms. Key challenges include limited linguistic diversity in corpora, restricted access to community-based local wisdom, and insufficient tools for managing vocabulary data effectively. To address these issues, the study suggests adopting corpus linguistic technologies to enhance data collection and analysis, fostering collaborations among academics and local communities, and developing vocabulary corpora that are applicable locally and globally. These efforts contribute to the creation of a standardized English vocabulary database, facilitating the preservation and international dissemination of health-related local wisdom while supporting academic and intercultural communication.

Keywords: challenges and applications; corpus linguistics; health-related local wisdom; vocabulary

Introduction

The rapid advancement of technology has facilitated seamless global communication, with English playing a critical role as the universal lingua franca for non-native speakers. Proficiency in English enhances opportunities for individuals in education, trade, employment, and communication. However, despite the recognition of English as an essential skill in Thailand, overall English proficiency remains below expectations. This shortfall highlights the ongoing challenge of improving English skills among Thai learners.

*Received: December 19 2024; Revised: December 25 2024; Accepted: December 26 2024

In academic and professional settings, the mastery of specialized vocabulary is crucial for effective communication and knowledge dissemination. Research indicates that academic texts often consist of 20-30% subject-specific terms (Fillmore, 1992), which are essential for understanding and producing scholarly work. However, Thai learners face significant challenges in acquiring such specialized vocabulary, as noted by Laufer (1989), due to limitations in access to structured and comprehensive vocabulary resources.

The preservation and dissemination of traditional health knowledge have gained increasing global attention for their contribution to sustainable healthcare practices. Van Rooy and Schäfer (2003) emphasize the necessity of systematically documenting traditional knowledge to ensure its accessibility and usability. Despite this, researchers face obstacles in creating standardized English terminology for traditional health knowledge, particularly when translating cultural-specific terms across languages and contexts. Hasko (2012) highlights that limited access to diverse linguistic data further compounds this issue, making it challenging to develop specialized vocabulary resources.

Linguistic corpora have become indispensable tools for addressing such challenges. The development of large-scale corpora, such as the British National Corpus (100 million words), the American National Corpus (22 million words), and the Hellenic National Corpus (over 34 million words), has significantly advanced language research and applications. These corpora support the creation of dictionaries, grammar references, and language teaching materials, enabling researchers to analyze real-world language usage comprehensively. In Thailand, however, efforts to develop linguistic corpora have been constrained by copyright issues and limited diversity in language data. For example, the ORCHID corpus by NECTEC is publicly accessible but restricted to academic texts, while other initiatives face challenges in ensuring comprehensive coverage.

The need for a specialized English vocabulary list specific to Thai traditional health knowledge is evident. Such a resource would not only standardize terminology for academic and professional use but also facilitate the global dissemination of this valuable cultural heritage. This study aims to address this gap by answering the research question: "How can a standard English vocabulary list for Thai traditional health knowledge be developed to enhance international academic communication?"

This research adopts a corpus-based approach to systematically identify and organize relevant vocabulary. By building on established linguistic theories and leveraging technological tools, the study seeks to create a robust vocabulary resource that aligns with academic and cultural preservation goals. Specifically, the objectives are twofold: to investigate the challenges in compiling such a vocabulary list and to explore effective methodologies for its development.

The outcomes of this study are expected to provide significant benefits. First, the standardized vocabulary list will enhance clarity and consistency in academic communication. Second, it will improve the accuracy of translations, allowing international audiences to better understand Thai traditional health practices. Finally, the study will serve as a foundational resource for future research, supporting both the preservation of traditional knowledge and the advancement of specialized vocabulary development. These contributions align with the broader mission of academic institutions to foster local knowledge and facilitate global engagement.

Research Objectives

1. To investigate the challenges in applying corpus linguistics for compiling a vocabulary list related to local wisdom in health contexts.

2. To explore approaches for applying corpus linguistics to develop a vocabulary list of health-related local wisdom.

Research Methodology

This study employs qualitative research, utilizing the documentary research method as a primary tool. The researcher selected secondary documents, including relevant literature that aligns with the research objectives. These consist of research articles and academic publications. The data collection process involved reviewing academic articles and research studies related to local health wisdom, theoretical frameworks, and conceptual understandings of local knowledge. The vocabulary identified and cited in these studies was drawn from 30 research articles within the Thai Journal Citation Index (TCI) database.

Research Tools

The tools used in this study include WordSmith Tools and AntWordProfiler, which analyze word frequency and lexical profiles.

Data Collection Methods

The researcher selected data from academic articles and research papers related to local health wisdom. Files in PDF format were converted into text files (.txt) before being processed with WordSmith Tools and AntWordProfiler. Data collection involved purposive sampling, selecting academic articles that focused on local health wisdom published online.

Data Analysis

The data analysis in this study was conducted using secondary documents, following the criteria proposed by Scott (1990; 2006). These criteria ensure the reliability and relevance of the documents selected for the research. The key principles include:

1. Authenticity: The documents must originate from credible and reliable sources, ensuring their accuracy, completeness, and alignment with the context in which they were published.

2. Credibility: The documents must be free from errors or any form of data distortion.

3. Representativeness: The documents must serve as representative samples of similar types of materials and provide details that can reflect the characteristics of a broader group or population.

4. Meaning: The documents must be clear, comprehensible, and consistent with the research objectives, reflecting their significance for the study (Mogalakwe, 2006).

Once these criteria were applied to the selected materials, the data were analyzed through a two-step process: data preparation and word frequency analysis.

In the first step—data preparation—the researcher selected research articles from the Thai Journal Citation Index (TCI) using purposive sampling. All selected articles were in PDF format, which were then converted into text files (.txt) to enable processing through word frequency analysis tools. Since the software programs used in the study—WordSmith Tool and AntWordProfiler—only accept text file formats, this conversion was necessary. During this stage, irrelevant content, such as Thai-language text, bibliographic references, appendices, and unrelated sections, were removed to ensure that only relevant content about health-related local wisdom remained in the text files.

The second step involved analyzing the word frequency of the prepared data using WordSmith Tool and AntWordProfiler. These programs were employed to identify and rank the most frequently occurring words in the selected articles. The results provided a clear overview of word occurrences, allowing the researcher to identify relevant vocabulary for the development of an academic word list related to health-related local wisdom.

This systematic approach ensured that the data analysis process was rigorous and aligned with the research objectives, providing a foundation for compiling a reliable and meaningful vocabulary list.

Conceptual Framework for the Study

The conceptual framework for this study explores the challenges and approaches for utilizing corpus linguistics in compiling vocabulary related to local health wisdom. The study focuses on identifying key issues and pathways for compiling a vocabulary list through corpus linguistic methodologies. This involves collecting and organizing vocabulary that aligns with local wisdom in health contexts.

The framework includes two primary areas: analyzing challenges and exploring applications of corpus linguistics. The study employs keywords analysis and concordance analysis to compile vocabulary that reflects the context of local health wisdom. The anticipated outcome is the creation of a structured vocabulary list that can contribute to linguistic studies and local knowledge research. The framework can be summarized as follows:

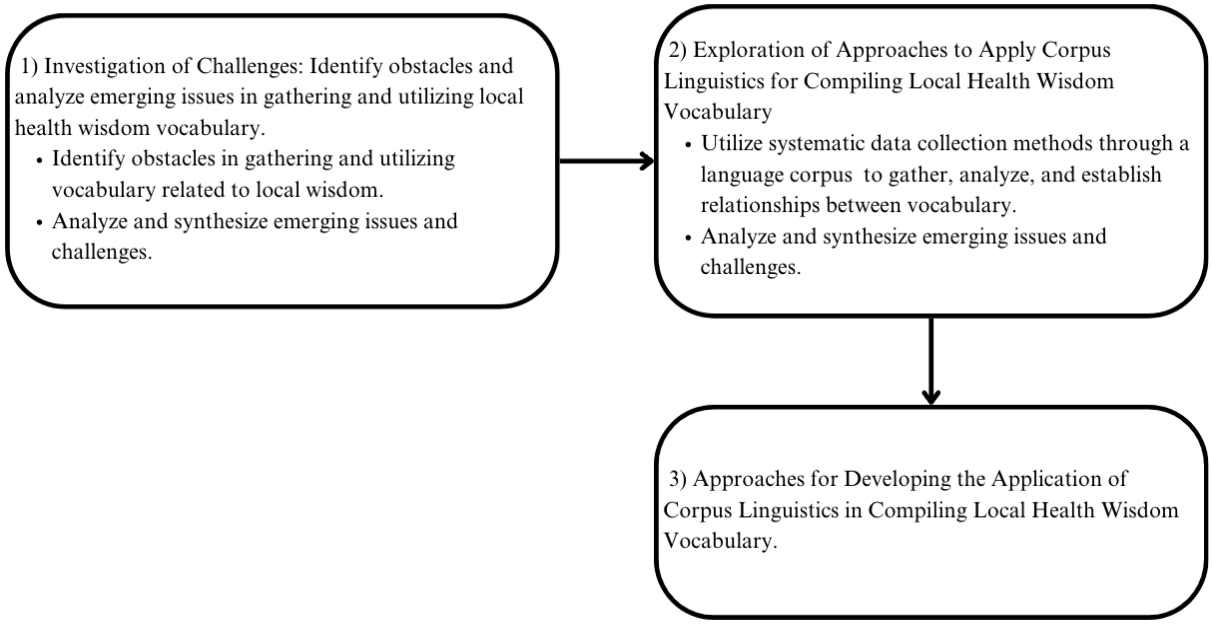

Figure 1. Conceptual Framework for the Study

Research Findings

This study investigates the challenges and applications of corpus linguistics in compiling a vocabulary list for health-related local wisdom. The research focuses on analyzing 30 research articles and academic papers selected from the Thai Journal Citation Index (TCI) database. The process of vocabulary compilation was guided by frequency and lexical profiling criteria, using tools such as WordSmith Tool and AntWordProfiler to analyze the collected data. The findings are summarized as follows:

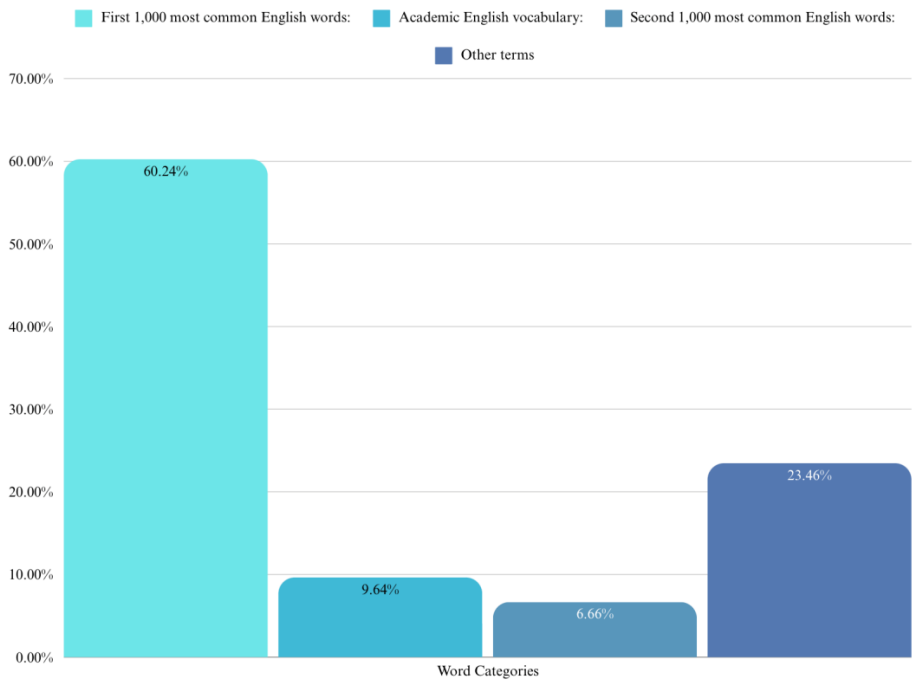

Figure 2. Proportion of word categories in the English corpus of health-related local wisdom.

As illustrated in Figure 2, the English corpus of health-related local wisdom vocabulary revealed that the largest proportion of words belonged to the General Service List (GSL), specifically the first 1,000 most common English words, accounting for 60.24% or 71,154 words. This was followed by academic vocabulary from the Academic Word List (AWL), making up 9.64% or 11,381 words. The third group consisted of words from the second 1,000 most common English words, at 6.66% or 7,864 words, while other words accounted for 23.46% or 27,711 words. After collecting and organizing the data, the next step involved filtering and selecting appropriate words to include in the specialized vocabulary list.

Despite these efforts, the findings highlight several challenges. First, linguistic diversity within the corpora remains limited, as most articles used in the study were domestic research papers written primarily in Thai, with only the abstracts available in English. Consequently, specialized vocabulary was underrepresented, and most of the identified words were common or academic English terms. Additionally, community-based knowledge often lacks standardized terminology, and existing tools for collecting and managing vocabulary data were found to be insufficient. These limitations necessitated modifications to the research process, including revising search terms and consulting international databases to enhance the variety and comprehensiveness of the vocabulary collected.

To address these challenges, the study implemented several approaches. By applying corpus linguistics methodologies, such as frequency analysis and lexical profiling, the research ensured the systematic selection of context-relevant terms. Further efforts included broadening the scope of data collection through targeted searches in academic books, international research papers, and articles related to health-related local wisdom. These methods resulted in a more diverse vocabulary collection and contributed to the foundation of a specialized English vocabulary list.

In summary, this study explored the application of corpus linguistics tools to compile a specialized vocabulary list for health-related local wisdom, addressing challenges associated with linguistic diversity and limited access to standardized terminology. The findings provide valuable insights into the composition of English vocabulary in this context and suggest practical approaches for improving vocabulary compilation. While the outcomes highlight the potential for enhancing the systematic development of specialized vocabulary lists, further work is needed to refine data collection methods and expand access to diverse linguistic resources to fully support the preservation and academic dissemination of health-related local wisdom.

Discussion

This study provides an in-depth analysis of the challenges and strategies for compiling a specialized vocabulary list related to health-related local wisdom using corpus linguistics. The discussion is structured to address the research objectives and to compare the findings with existing literature, providing a nuanced perspective on the challenges encountered and solutions proposed.

The research identified three major challenges in developing a vocabulary list. Firstly, linguistic diversity within the collected data was limited. The findings revealed that the highest proportion of terms belonged to the General Service List (GSL), with 60.24% or 71,154 words. This was followed by academic vocabulary from the Academic Word List (AWL), accounting for 9.64% or 11,381 words. Words from the second 1,000 most common English words made up 6.66% or 7,864 words, while 23.46% or 27,711 words fell into other categories. These results align with Tono (2003), who emphasized that well-designed corpora must provide clear criteria for data collection, as this ensures their utility for linguistic analysis. However, this study highlights a unique challenge not previously addressed: the over-reliance on English abstracts in Thai research articles, which limits the scope of specialized vocabulary.

Secondly, the study uncovered challenges related to the accessibility of community-based knowledge. Specifically, this knowledge often lacks standardized terminology, which restricts its integration into academic and global discourse. This observation is consistent with Van Rooy and Schäfer (2003), who noted that non-native linguistic tools may not fully capture the specificity of specialized knowledge domains. Furthermore, insufficient tools for compiling and managing vocabulary data posed a significant barrier. The reliance on tools developed for native English corpora, such as WordSmith and AntWordProfiler, highlights the need for localized adaptations to accommodate the unique characteristics of non-native datasets.

To address these challenges, this study implemented several strategies. For instance, corpus linguistics methodologies, including frequency analysis and lexical profiling, were applied to systematically filter and select terms. While these methods proved effective, the study expanded data collection by revising search terms and incorporating international datasets. This approach aligns with Hasko (2012), who advocated for the inclusion of global resources to enhance linguistic diversity and comprehensiveness. By broadening the scope, this study contributed to a more robust and contextually relevant vocabulary list.

The findings and approaches of this study also align with previous literature in several ways. For instance, the emphasis on word frequency and profiling corroborates Laufer's (1989) research, which highlighted the importance of lexical profiling in identifying domain-specific terms. However, unlike Laufer, this study focuses exclusively on health-related local wisdom, thus filling a critical gap in the field. Additionally, the integration of international datasets to address linguistic diversity extends Hasko's (2012) recommendations, demonstrating their practical application in a specific research context.

In conclusion, this study highlights the importance of integrating corpus linguistics tools and collaborative strategies in compiling specialized vocabulary lists. By supporting processes such as data collection, analysis, and vocabulary linkage, this research underscores the need for interdisciplinary collaboration between academic researchers and community stakeholders. Such initiatives are essential for fostering the development of functional vocabularies at both local and global levels. Furthermore, the findings contribute to the sustainable transmission and global recognition of health-related local wisdom while promoting a deeper understanding of its value in contemporary contexts. These findings serve as a foundation for future research aimed at expanding and refining specialized linguistic resources.

Recommendations

1. Academic Suggestions

- Researchers should adopt broader and more inclusive topics to address the limitations caused by the predominance of Thai-language articles with limited English content, ensuring access to diverse and comprehensive datasets.

- Expanding the scope of research topics will enhance the range and utility of future vocabulary development projects.

2. Policy Suggestions

- Encourage collaboration among researchers, academics, and local communities to develop specialized vocabulary corpora and establish domain-specific vocabulary databases.

- Promote integrated approaches to foster partnerships across diverse fields of expertise to support sustainable vocabulary development efforts.

3. Practical Suggestions

- Involve stakeholders in shaping and implementing strategies for developing linguistic databases, ensuring the systematic compilation of health-related local wisdom vocabulary lists.

- Implement continuous monitoring and evaluation mechanisms to ensure the effectiveness and relevance of policy implementation and the resulting domain-specific vocabulary databases.

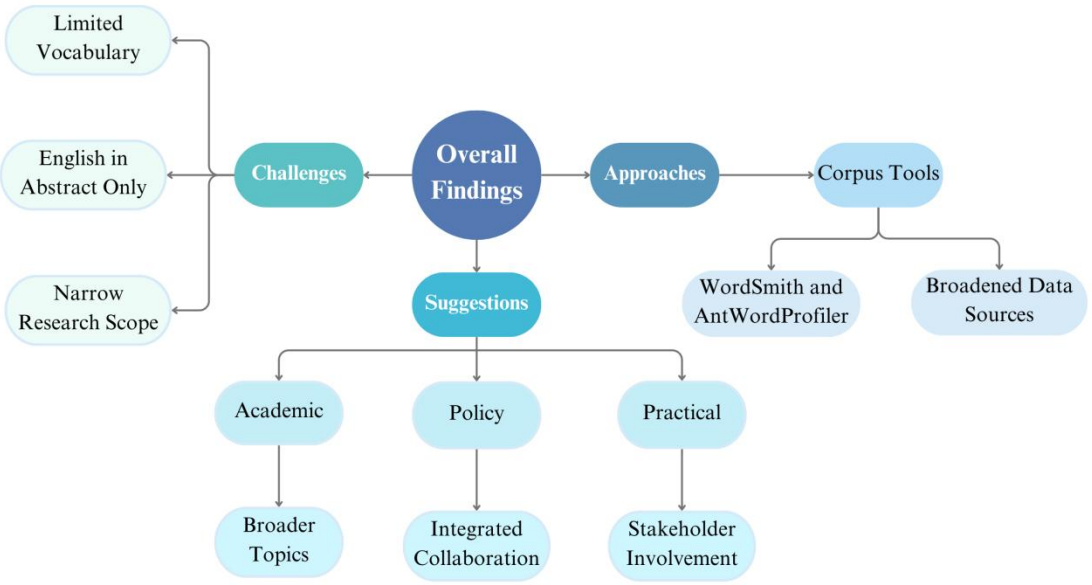

Figure 3. Mind map synthesizing the study's findings and categorized suggestions for developing a vocabulary list related to health-related local wisdom.

References

Baker, W. (2003). Should culture be an overt component of EFL instruction outside of English-speaking countries? *The Thai context*. *Asian EFL Journal, 5* (4). Retrieved April 30, 2011, from http://www.asian-efl-journal.com/dec_03_sub.wb.php

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34* (2), 213–238.

Fillmore, C. (1992). Corpus linguistics or computer-aided armchair linguistics. In J. Svartvik (Ed.), *Directions in corpus linguistics: Proceedings of Nobel Symposium 82* (pp. 35–60). Berlin: Germany: Mouton de Gruyter.

Foley, J. A. (2005). English in Thailand. *RELC Journal, 36* (2), 223–234.

Hasko, V. (2020). Qualitative corpus analysis. *The Encyclopedia of Applied Linguistics*, 1–7. https://doi.org/10.1002/9781405198431.wbeal0974.pub2

Laufer, B. (1989). A factor of difficulty in vocabulary learning: Deceptive transparency. *AILA Review, 6*, 10–20.

Laufer, B. (1992). Reading in a foreign language: How does L2 lexical knowledge interact with the reader's general academic ability? *Journal of Research in Reading, 15*, 95–103.

Mogalakwe, M. (2006). The use of document research methods in social research. *African Sociological Review, 10* (1), 221–230.

Saussure, F. (1966). *Course in general linguistics*. (C. Bally & A. Sechehaye, Eds.; W. Baskin, Trans.). New York, NY: McGraw-Hill.

Scott, J. (1990). *A matter of record: Documentary sources in social research.* Cambridge: Polity Press.

Scott, J. (2006). Social research and documentary sources. In *SAGE benchmarks in social research methods: Documentary research* (Vol. 1, pp. 3–40). SAGE Publications.

Tono, Y. (2003). Learner corpora: Design, development and applications. In P. R. D. Archer, A. Wilson, & T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 conference* (pp. 800–809). Lancaster University.

Van Rooy, B., & Schäfer, L. (2003). Automatic POS tagging of a learner corpus: The influence of learner error on tagger accuracy. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 conference* (pp. 835–844). Lancaster University.