PASAA Journal
Volume 71, July – December 2025, 269–306

PASAA

A Baking Verb List: A Corpus-Based Study of International Bakery and Pastry Recipes

Woravit Kitjaroenpaiboon[a]*, Sirote Pholpuntin[b], Sukhum Chaleysub[c], Samniang Fahkrajang[d], Prissana Fongsrun[e], Niphatchanok Najpinij[f], and Chanchana Siripanwattana[g]

[a,b,c,d,e,f,g] Suan Dusit University, Bangkok, Thailand

**Corresponding author: woravit_kit@dusit.ac.th*

Article information	
Abstract	The research applies corpus linguistics methods to discover essential words in the Suan Dusit University International Bakery and Pastry Corpus (SDUIBPC). The SDUIBPC contains 5,484,999 words which were collected from 100 English cookbooks that focus only on baking recipes. The research investigates words in international bakery recipes through a focus on verbs which define this specific discourse. The research combines AntConc 4.3.1 with statistical methods including frequency and range analyses, log-likelihood, Bayes factor, effect size for log-likelihood and evaluations from experts. The research discovered 761 statistically significant words which distribute across four functional categories including nouns, verbs, adjectives and adverbs. The functional analysis revealed 103 common verbs which demonstrate the operational and performative characteristics of culinary language. The research results provide essential language knowledge which helps culinary experts, teachers, and students enhance their English skills for specialized environments. The

E-ISSN: 2287-0024

	research establishes a path for culinary students to learn modern professional culinary terminology.
Keywords	lexical verbs, international bakery recipes, linguistic corpus, corpus linguistics
APA citation:	Kitjaroenpaiboon, W., Pholpuntin, S., Chaleysub, S., Fahkrajang, S., Fongsrun, P., Najpinij, N., & Siripanwattana, C. (2025). A baking verb list: A corpus-based study of international bakery and pastry recipes. *PASAA Journal*, *71*, 269–306.

1. Introduction

The bakery industry has experienced major changes because of the COVID-19 pandemic, which started in 2019. Consumers needed new solutions because of social distancing rules, lockdowns, and changing consumer habits (Cheevanon, 2022). The bakery industry faced two contrasting trends: some businesses shut down while others gained more customers (Smanalieva, 2025). The lockdown period brought about a major increase in online bakery sales which demonstrated how consumer behavior evolved (Żurek & Rudy, 2024).

Bakeries now focus on offering healthier options because customers want products with gluten-free, plant-based, organic, and clear nutritional information (Sonia et al., 2019). People choose homemade baked goods made in their local area because they believe the products have better taste and quality (Dessev et al., 2020). Bakeries focus on sustainability through their adoption of eco-friendly manufacturing methods, biodegradable materials, local ingredient sourcing, and waste reduction initiatives. The global bakery industry combines different flavors in its products because customers want to experience new and exciting culinary tastes (Martínez-Monzó et al., 2013).

Baked goods hold essential value in all societies because they function as fundamental food items throughout different cultures (Dessev et al., 2020). The

development of baked goods depends on geographical conditions, economic factors, and climate as well as cultural elements (Chatterjee et al., 2016). Baked goods maintain a deep connection to cultural traditions because they serve as essential components during special events (Wesser, 2021). Bakeries function as community centers which enable people to meet while supporting economic growth through employment opportunities (Azanedo et al., 2020). The baking industry drives culinary progress through continuous developments in flavors, ingredients, and preparations (Sutton, 2017). The bakery industry creates effects that reach beyond its boundaries to impact the entire food industry. This demonstrates its vital role in nutrition, culture, economy, sustainability, and innovation (Rokach, 2020).

People now travel to different locations for bakery tourism because bakery consumption has become more popular worldwide (Żurek & Rudy, 2024). The growing number of new bakeries in Thailand has become more visible, particularly in media outlets (Cheevanon, 2022).

Baked goods appeal to everyone, but language differences create barriers which may inhibit people from sharing their culinary traditions. English functions as a worldwide communication system, but not everyone masters its use (Moussu & Llurda, 2008). The bakery industry in Thailand faces English communication challenges because English serves as a foreign language in the country (Ambele, 2022). The development of bakery-specific English skills would enable Thailand to participate in international bakery traditions while preserving its native baking heritage. This linguistic improvement could help promote Thai bakery worldwide, enhance Thai soft power, and draw international attention toward Thailand's food heritage.

The research addresses these challenges through the development of the Suan Dusit University International Bakery and Pastry Corpus (SDUIBPC) which contains 5,484,999 words from 100 English-language international bakery recipe

books. The corpus contains authentic bakery language to support academic studies and educational purposes. The research particularly investigates verbs because they serve as essential words to describe baking preparation procedures. The main goal of this research thus involves identifying the most common baking verbs which appear in the SDUIBPC.

2. Literature Review

For the purpose of this research, the relevant literature on English corpora for specific purposes will be structured into three parts. The first section, Concepts of Corpus, discusses the principles and significance of corpora in linguistic research. This section emphasizes their role as databases for lexical analysis. The second section, Corpus Analysis, explores the methodologies used to examine corpora. The section aims to understand linguistic patterns and structures. The third section, Lexical Analysis in the Corpus, focuses on the identification and interpretation of vocabulary within corpora. This provides insights into word meanings and usage.

2.1 Concepts of Corpus

A corpus is an extensive collection of written or spoken language data which is stored in a computer system for linguistic analysis. Typically stored as plain text, corpora can be analyzed with concordance software to search for words, calculate frequency and range, and display results in Key Word in Context format. This facilitates the study of word usage and collocations in contexts.

The SDUIBPC is a specialized, monolingual, written corpus focusing particularly on English used in the bakery domain. It was compiled by the research team at Suan Dusit University from a collection of 100 international bakery recipe books and contains a total of 5,484,999 words. The corpus is designed to provide a representative sample of authentic lexical usage in the international baking discourse and serves as a resource for linguistic analysis, language learning, and culinary research.

E-ISSN: 2287-0024

Research in linguistics benefits from corpus analysis tools known as concordancers, which help researchers detect word occurrences and their relationships. The research tools AntConc (Anthony, 2024), WordSmith (Scot, 2024), and MonoConc (Barlow, 2000) represent popular software options.

The analysis of corpora enables researchers to improve understanding of vocabulary and accuracy of translation through direct linguistic evidence. The research data shows which words and expressions appear most frequently in natural contexts so researchers can select the most important terms for their studies. This evidence-based knowledge also helps teachers to pick suitable vocabulary for instruction, translators to achieve better results, and students to learn word applications in actual language use (Endoo, 2017). Moreover, corpora help translators select appropriate words, idiomatic expressions, and register-specific language which results in clearer, more natural, and more accurate translations. Bilingual or parallel corpora enable translators to view actual language usage between source and target languages through aligned text pairs, which helps them create more contextually accurate and relevant translations (Zanettin, 2014).

According to Conrad (1999), three essential elements which affect corpus analysis include:

1. **Corpus Size** – the reliability of corpus analysis depends on the total number of texts included in the corpus. The required corpus size for research depends on the specific goals of the study. Research on domain-specific vocabulary and linguistic phenomena can use small specialized corpora, but studies that need general linguistic patterns require larger ones. The size of corpora has no predefined boundaries since researchers determine the appropriate amount based on their study goals.
2. **Computational Analysis** – the process of data collection and analysis depends heavily on computers because they enable researchers to

generate frequency counts and concordances. This study used AntConc 4.3.1 (Anthony, 2024), which helped the researchers to better understand the data, and to analyze word frequencies and ranges in the SDUIBPC.

3. **Quantitative and Qualitative Approaches** – the analysis of corpora requires researchers to use quantitative methods for frequency analysis together with qualitative methods for semantic analysis and evaluations.

Sinclair (2014) explains that corpus linguistics operates differently from traditional linguistic research because it analyzes actual language usage instead of established grammatical rules. Halliday (1992) developed the lexicogrammar theory which demonstrates how words and grammatical structures exist in a connected system. Halliday's theory is about grammatical structures, regular linguistic patterns, and how words affect syntax in real-world texts. An analysis of a corpus helps reveal actual language patterns which occur in both spoken and written communication. The combination of quantitative frequency and range analyses with qualitative semantic categorization provides a complete linguistic understanding.

2.2 Vocabulary List Development

The creation of English teaching vocabulary lists began in the early 20th century, and corpus analysis has since become a key factor in their development. The advancement of new technology and computational linguistics methods enables faster and more effective vocabulary compilation (Babazade, 2024).

Ostonova and Xikmatovna (2020) note that the first English vocabulary lists appeared during the 16th century. The creators of these early glossaries remain unknown, but scholars and educators compiled them for teaching purposes to focus on fundamental words about animals, body parts, and occupations. The early lists did not mention any criteria for selecting words to be included. Then, in the 19th century, linguists understood the need for fundamental vocabulary. Therefore, they created more systematic vocabulary lists, which included

Häufigkeitswörterbuch der deutschen Sprache as the first frequency-based vocabulary list. The list contains basic words from everyday conversations to help students learn language and to support lexicographers (Jones & Tschirner, 2015).

In the 20th century, vocabulary lists were developed for practical purposes, including education and language learning. However, Thorndike's widely used vocabulary list, compiled in the 1920s, introduced the semantic count technique, emphasizing frequency, range, and cross-referencing with other frequency lists (Gilner & Morales, 2010). Thorndike's list for students' vocabulary development affected educational practices, but several experts criticized it since the list failed to recognize the differences between polysemous words and idiomatic expressions. Palmer (1938) thus created the core vocabulary list which included the headword system that later vocabulary lists adopted his idea (Nation, 2016).

In 1953, West (1953) developed the General Service List (GSL) through his analysis of a 5-million-word corpus. West's core purpose was to help language learners develop basic communication skills through his original list of high-frequency English words (Gilner & Morales, 2010). West's GSL has become popular, but experts criticize its outdated word selections and limited inclusion of contemporary specialized terms (Nation & Kyongho, 1995).

In 1984, the University Word List (UWL) by Xue and Nation (1984) became a fundamental reference for language research. The list was criticized as having several drawbacks because it did not include enough academic subjects, used outdated academic materials, and included words that appeared infrequently in learning materials.

In 2000, the Academic Word List (AWL) by Coxhead addressed previous list limitations and has become a standard reference for language education (Nation, 2016). Nevertheless, the AWL has two main drawbacks because it presents legal and economic terms unevenly and only includes written academic materials

(Ostonova & Xikmatovna, 2020). Consequently, the development of new academic word lists has continued through the work of Gardner and Davies in 2014 and Dang, Coxhead, and Webb in 2017.

The AWL by Coxhead has led to the creation of specific word lists which serve different academic fields including applied linguistics, humanities, business management, science, medicine, and engineering. These lists have been developed using corpus approaches based on methodologies by West in 1953 and Coxhead in 2000 (Nation, 2016). For example, corpus-based studies have identified at least two medical vocabulary lists (e.g., Nguyen & Miller, 2020; Wang et al., 2008), two engineering lists (e.g., Puangmali, 1976; Ward, 2009), and three food service vocabulary lists (e.g., Kitjaroenpaiboon et al., 2024; Nordin et al., 2013; Rungrueang et al., 2022). The above examples show how frequency-based lexical analysis produces specialized vocabularies for different fields of study.

As can be seen, the historical development of vocabulary lists from the 16th century until today shows how frequency and range have become essential selection criteria. The term *range* in this context refers to the number of texts in a corpus which contain a word, indicating how frequently the word appears across different texts.

2.3 Food and Bakery Vocabulary List Development

In 2013, Nordin et al. (2013) developed a specific vocabulary list for culinary writing education. The researchers built their 3,698-word corpus from materials which their students learned during culinary writing classes. The researchers used lexical frequency and range analysis to identify specialized vocabulary. They then developed a food terminology list containing 112 specific terms.

In 2022, Rungrueang et al. (2022) worked with their team to improve food service industry staff's lexical abilities by developing industry-specific vocabulary which matched the staff's work environment. The Food Service Corpus (FSC),

which included 1.8 million words from four food websites, served as the key source for this list. The analytical process required three essential stages which started with keyword identification, followed by statistical analysis, through log-likelihood measures, and ended with expert evaluation from three food specialists who validated 261 food service-specific terms.

In 2024, Kitjaroenpaiboon et al. (2024) studied the most common words in the Suan Dusit University International Recipe Corpus (SDUIRC), which contains more than 7 million words. They analyzed the corpus using AntConc 4.3.1 (Anthony, 2024) and applied frequency and range analysis, log-likelihood, Bayes factors, and effect sizes for log-likelihood, along with evaluations from experts. Their study identified a total of 1,165 frequently occurring words in the corpus.

In conclusion, although previous studies have investigated specialized vocabulary in food-related contexts using corpus-based analysis, no research to date has focused on identifying frequently occurring English verbs in any corpus of international bakery recipes. While the work of Kitjaroenpaiboon et al. (2024) represents an important contribution to the study of vocabulary in an international food recipe corpus, it addressed recipes in general and did not isolate the distinct domain of bakery recipes.

3. Methodology

This research was granted ethical approval under the certification number SDU-RDI-SHS 2024-093. The research depended mainly on a corpus as its fundamental element. The researchers generated the SDUIBPC to analyze international bakery recipes through systematic vocabulary identification. The researchers collected 5,484,999 words from 100 international bakery recipe books to create the corpus. The collection includes a variety of popular bakery recipes including bread, pastries, cakes, and desserts, which makes the SDUIBPC a suitable representation of actual English baking texts.

The research team of this study chose texts based on their ability to support generalizable results that represent international baking recipe language patterns. The academic consensus supports that specialized corpora need at least 20,000 words to perform a meaningful analysis (Gries, 2009). However, the SDUIBPC contains 5,484,999 words, which exceeds the word count of previous studies including Can et al. (2016) and Rungrueang et al. (2022). According to Gries (2009), the SDUIBPC meets corpus research standards for a complete lexical analysis since it contains more than enough words.

The research team obtained the SDUIBPC from the 100 most popular English baking recipe books available on https://www.pdfdrive.com, which operates as a digital library that contains more than 85 million books from different categories. The research team used these books only for academic purposes while avoiding any copyright violations through content reproduction or distribution. The research team always adhered to copyright and fair-use guidelines, ensuring ethical use of these materials for corpus compilation.

As mentioned earlier, the selection of the 100 books in this study produced a bigger and more diverse collection of data than previous studies, which enhances the reliability of vocabulary analysis. This research employed the following selection criteria:

1. The research team used corpus linguistics methods which Baoya (2015), Getkham (2014), Kitjaroenpaiboon et al. (2021a, 2021b), and Kitjaroenpaiboon and Getkham (2015, 2016, 2017) employed to select English-language international bakery recipe books based on their ranking among the top 100 most downloaded texts in that genre. This study used this method to obtain typical and important language patterns from English-language international bakery recipe books.
2. The research team applied the corpus linguistics guidelines of Kitjaroenpaiboon and Getkham (2015, 2016, 2017) to handle situations where different versions of the same book existed with different cover

designs or publication dates. The researchers only selected one version of each book for analysis while adding the next book from the ranking list to achieve their goal of 100 unique and different books.

3. The researchers discovered that particular recipe books contained three separate sections, which included bakery recipes, savory dishes, and beverage preparation. This research aimed to investigate only the lexical elements which appeared particularly in international bakery recipes. The researchers selected baking-focused books only and excluded publications that combined baking content with savory dishes and beverages.
4. This study did not establish any distinction between books written by native English speakers and books written by non-native English speakers. The researchers state that attempting to determine author native language through name inspection proved both difficult and unreliable. The publication process for English-language international bakery recipe books needed to include multiple stages of quality evaluation and language editing and editorial review. Thus, the books were able to serve as model examples which demonstrated how the English language appeared in bakery recipe book creation.

The researchers converted the 100 English-language international bakery recipe book files from https://www.pdfdrive.com into .txt format to create the SDUIBPC for the lexical analysis and vocabulary compilation. The research team performed required edits and removals of specific content in the recipe books in the SDUIBPC to verify the accuracy of vocabulary analysis. The researchers applied corpus analyst guidelines (Baoya, 2015; Getkham, 2010; Kitjaroenpaiboon et al., 2021a, 2021b; Kitjaroenpaiboon & Getkham, 2015, 2016, 2017) to make the following adjustments:

1. Sections, including the cover, preface, table of contents, acknowledgments, keywords, bibliography, index, and author biography,

were removed, as these parts did not contribute to the analysis of baking vocabulary and may have led to errors in frequency analysis.

2. Any spelling or spacing errors found in the content were corrected to prevent inaccuracies in the word frequency analysis.
3. Any formulas or special characters were removed to avoid errors in the compilation of frequently occurring words in the corpus.
4. Citations within parentheses were deleted, as they could have caused errors in the word frequency analysis.

The research methodology required a pilot study after finishing the revisions.

The SDUIBPC included only baking recipes for its selection process. The researchers discovered that English-language international bakery recipe books contained multiple food categories including baked goods, savory dishes, and beverages. The researchers sought expert opinions from four international baking specialists to establish food category definitions for all recipe books because they wanted to achieve precise, dependable results, and prevent personal interpretation. This method followed Cargill and O'Connor (2009) who recommend using three experts with appropriate knowledge to reduce classification errors.

It was necessary for the food classification analysis within the corpus to maintain both reliability and accuracy. According to Gwet (2014) and Neuendorf (2002), such analysis requires human evaluation which should pass an intercoder reliability test. The analysis required three participants to use the same framework for result comparison. The analysis results were statistically combined through methods described by Sicanore et al. (1999). The researchers and four baking experts collaborated to create a classification system which separated recipes for baked goods from those for savory dishes. The definition of baked goods matched that of Riquelme et al., (2022) who explains that baked goods require oven or heat source baking to achieve their cooked state. Baked goods include cookies, cake,

bread, muffins, and pies which may require ingredients such as eggs, flour, sugar, milk, butter, yeast, and baking powder for preparation.

The researchers together with the four experts started the analysis of the first 50 books from the corpus based on the established classification framework. The experts used this first stage to learn about classification methods while they detected any problems that needed clarification. The researchers helped handle any problems which emerged during this stage. The participants reached a shared understanding about definitions and classifications through their initial work together. The experts then conducted individual analyses of the remaining 50 books after completing the familiarization process while using the established classification framework. Neuendorf (2002) suggests that an agreement rate of 80% or higher may be considered sufficient for analysis results. The researchers performed additional analysis when the agreement rate fell below 80% and continued until they achieved the required agreement rate of 80% or higher.

The researchers used AntConc 4.3.1 (Anthony, 2024), MS Word 2013, and https://th.wordcounter360.com to determine the word count. The three programs produced identical word count results, which demonstrated their ability to perform accurate and reliable word counting (reliability value = 100%). The SDUIBPC contains 100 bakery recipe books which together contain 5,484,999 words. The books in the collection contained an average of 54,850 words each (average = 54,850). The English-language international bakery recipe book that contained the most words had 315,066 words while the book with the fewest words contained 3,325 words. The standard deviation was 56,108.75, indicating a substantial variation in the length of recipe books in the corpus, with some books containing many more words than others. This variability reflects differences in the comprehensiveness, scope, and content density of the sources included in the SDUIBPC.

After the researchers and four experts analyzed the food categories in the corpus using Riquelme et al.'s (2022) food classification concept and applied the concept of acceptable consensus analysis results proposed by scholars (e.g., Cargill & O'Connor, 2009), the next step was to identify frequently occurring vocabulary in the SDUIBPC.

The analysis and compilation of frequent words or vocabulary in the SDUIBPC were conducted by using AntConc 4.3.1 (Anthony, 2024). In this study, the term *vocabulary* is defined according to Coxhead (2000) as words that frequently appear with a high usage frequency in the corpus. Corpus linguists (e.g., Biber et al., 2002; Qurboniyozovna, 2025) suggest that words considered as frequent should appear at least five times per 100,000 words in the corpus and occur in at least five distinct texts. This study used words which appeared frequently in different texts to achieve this goal. The SDUIBPC corpus contains 5,484,999 words in total. A word in the corpus needed to appear at least 275 times (frequency = 275) and be present in five or more different texts (range = 5) to meet the requirements. The researchers used AntConc 4.3.1 (Anthony, 2024) to analyze and compile the most common words after defining the minimum rate of word occurrence in the SDUIBPC.

The researchers employed AntConc 4.3.1 (Anthony, 2024) and Monoconc (Barlow, 2000) to analyze, crosscheck, and identify the most common words from the selected international bakery and pastry recipe books. The researchers recognized that running an analysis through a secondary corpus tool helped confirm the accuracy of the analysis. The corpus analysis tool Monoconc (Barlow, 2000) enabled the researchers to study word occurrence patterns and distribution throughout multiple texts. The researchers again followed the recommendation from Gwet (2014) and Neuendorf (2002) to verify analysis reliability. As aforementioned, the researchers employed Monoconc (Barlow, 2000) as a supplementary tool to analyze and compile vocabulary. The analysis results from both tools produced identical vocabulary counts. The results indicated that

AntConc 4.3.1 (Anthony, 2024) and Monoconc (Barlow, 2000) produced identical vocabulary counts, which confirmed their reliability for vocabulary analysis and compilation. These results confirm Ari's (2006) finding that Antconc 4.3.1 (Anthony, 2024) and Monoconc (Barlow, 2000) successfully identify matching vocabulary sets.

Once the frequent vocabulary from the 100 English-language international bakery recipe books was compiled, the researchers and the four experts in bakery practices discussed and evaluated the importance of the identified vocabulary. The goal was to determine whether the vocabulary should be included in the final list of frequent terms.

The analysis used the Word and n-gram functions in AntConc 4.3.1 (Anthony, 2024) to identify vocabulary, rather than relying on the Keyword function. While the Keyword function is practical for identifying statistically significant single words, it cannot capture multi-word units such as noun phrases (e.g., baking powder, chocolate ganache), which are common in the SDUIBPC corpus. This made the n-gram function more appropriate. In AntConc, the term *n-gram* is used to describe a string of words that co-occur in a text. To capture single-word and multi-word units, this study used n-grams ranging from 1-grams (single-word) to 5-grams (five-word sequence).

The significance of vocabulary (Keyness) was evaluated using multiple complementary methods. The log-likelihood statistic was used to compare word frequency differences between the SDUIBPC and nine reference corpora through *p*-values, which were considered to have statistical significance at less than .05 (Dunning, 1993; Rayson et al., 2004).

The nine reference corpora consisted of: the Corpus of Contemporary American English (COCA), the Coronavirus Corpus, the Global Web-Based English Corpus (GloWbe), the Movie Corpus, the Corpus of American Soap Operas, the TV

Corpus, the Wikipedia Corpus, the Multidisciplinary Academic Corpus (Kitjaroenpaiboon et al., 2021b), and the Suan Dusit University International Recipe Corpus (SDUIRC) (Kitjaroenpaiboon et al., 2024). The nine selected corpora represented different genres, registers, and contexts including general English, academic, media, online, and specialized culinary texts for complete SDUIBPC vocabulary analysis against English language usage.

The analysis included the SDUIRC because this international recipe corpus shares thematic content with the SDUIBPC. The SDUIRC served as a crucial reference point because it allowed the researchers to differentiate between standard cooking terminology in recipes and the specific baking terminology found in the SDUIBPC. The addition of this corpus enabled the researchers to identify baking discourse lexical characteristics with greater accuracy against the background of general culinary terminology.

Moreover, the Bayes factor calculations were used to evaluate observed frequency data under two alternative hypotheses. The calculation enabled the researchers to determine the strength of word distinctiveness within the target corpus (Kass & Raftery, 1995). The effect size for log-likelihood was used to measure the size of the difference, which helped the researchers understand the practical impact of results beyond statistical significance. The expert evaluation process brought together two international baking experts and two native English-speaking faculty members who specialized in syntax to verify if statistically significant terms held practical value in baking situations. As can be seen, this research combined statistical and practical assessment methods to identify baking vocabulary having both statistical importance and practical value in actual baking situations.

Once the words were extracted from the SDUIBPC, their functional categories in the SDUIBPC were examined based on the frameworks of established scholars (e.g., Anward, 2000; Bisang, 2011; Evans, 2000; Heine &

Kuteva, 2012; Sim & Haspelmath, 2012). The scholars state that English vocabulary typically includes eight main word functions: noun, pronoun, verb, adjective, adverb, interjection, conjunction, and preposition. As mentioned, this study focused on verbs since verbs are central to conveying procedural actions, steps, and processes essential in baking discourse. Accordingly, to check the classification accuracy, the frequently appearing word functions were analyzed in the English bakery recipes. Consequently, all the words found in this research were analyzed using the Multidimensional Analysis Tagger 1.3.3 program (Nini, 2021), in collaboration with the researchers, two experts in international baking, and two native English-speaking teachers, all of whom had expertise in syntax. During the analysis, the researchers, baking experts, and English teachers discussed and determined the functions of all the words as they appeared in sentences in the corpus. They then compared the analysis results with those obtained from the Multidimensional Analysis Tagger 1.3.3 (Nini, 2021). The process was deemed complete when all parties agreed on the classification of each word.

4. Results

To achieve the results for the research objective, the researchers applied the framework of corpus linguistics scholars (e.g., Biber et al., 2002; Coxhead, 2000; Gilner & Morales, 2010; Qurboniyozovna, 2025), which suggests that vocabulary must appear at least five times per 100,000 words and in at least five different texts to be considered frequent in a corpus. This criterion guaranteed that this study identified vocabulary with the highest frequency of occurrence that was generally encountered in the corpus. The SDUIBPC contains a total of 5,484,999 words. Therefore, the frequency count of any given word needed be at least 275 (frequency = 275), and the word needed to appear in at least five different texts (range = 5).

Table 1

Results of the Lexical Analysis of 761 Terms Found in the SDUIBPC

	Frequency	Range
Number	761	761
Mean	2,151.71	62.50
Standard Deviation	3,957.86	22.97
Median	853.00	63.00
Mode	344.00	76.00
Variance	15,644,087.36	526.86
Range	38,963.00	96.00
Maximum	39,239.00	100.00
Minimum	276.00	5.00

The analysis of SDUIBPC through AntConc 4.3.1 (Anthony, 2024) revealed 761 common words in the SDUIBPC. The 761 words appeared throughout the corpus with an average of 2,151.71 occurrences (SD = 3,957.86) while their occurrence numbers spanned from 276 to 39,239 (range = 38,963). The distribution of word frequencies showed some skewness because the median frequency was 853, the mode was 344, and the variance was 15,644,087.36.

The 761 words were distributed across an average of 62.50 texts (SD = 22.97) with a median of 63 texts, and a mode of 76 texts. They occurred in 5 to 100 texts (range = 96; variance = 526.86).

The analysis of word significance used log-likelihood to check for statistical differences at $p < .05$ (Dunning, 1993), Bayes factor to determine evidence strength and effect size for log-likelihood to measure difference sizes. The quantitative findings underwent expert validation for the researchers to achieve proper contextual understanding.

The researchers needed to perform a search for the 761 words across nine additional language corpora before they could analyze the above statistical measures. The researchers conducted a search for these words across nine

additional language corpora including the COCA, the Coronavirus Corpus, the GloWbE, the Movie Corpus, the American Soap Operas Corpus, the TV Corpus, the Wikipedia Corpus, the multidisciplinary academic corpus (Kitjaroenpaiboon et al., 2021b), and the SDUIRC (Kitjaroenpaiboon et al., 2024) to evaluate their occurrence in different corpora. The search results showed how the SDUIBPC word frequencies compared to those found in the nine reference corpora.

The research analyzed the predicted frequency counts of 761 terms which produced these results:

The SDUIBPC contained 761 frequently used terms, which showed an average expected frequency of 689.91 and a standard deviation of 1,198.24. The median frequency was 278.35 while the mode frequency was 84.69, which indicated that several words appeared frequently, but most words appeared infrequently. The variance was 1,433,886.02, while the frequency range extended from 74.44 to 11,101.44 (range = 11,027.00). The statistical data showed that the corpus contained few dominant words which appeared frequently, but most words appeared rarely.

The analyzed terms in the COCA showed a mean expected frequency of 689.98, while the standard deviation reached 1,198.20, which indicated significant variation in word appearances. The median frequency was 278.35, while the mode was 84.69 which indicated that several words appeared frequently, but numerous words appeared only rarely. The variance was 1,433,801.86, while the frequency range extended from 74.44 to 11,101.44 (range = 11,027.00). The statistical data demonstrated that most words in the corpus appeared infrequently because a limited number of words appeared frequently throughout the text.

The analyzed terms in the Coronavirus Corpus showed a mean expected frequency of 212.89, while the standard deviation was 369.69, which indicated that word occurrences varied greatly. The median frequency was 85.88, but the mode

frequency was 26.13, which indicated that few words appeared frequently while most words appeared rarely. The variance was 136,494.08, while the frequency range extended from 22.97 to 3,425.25 (range = 3,402.28). The statistical data showed that the corpus contained few dominant words which appeared frequently, but most words appeared infrequently.

The GloWbE showed a mean expected frequency of 145.78 for analyzed terms. The word frequencies in the corpus showed significant variation because the standard deviation was 253.16. The median word frequency was 58.81, while the mode frequency stood at 17.89. The word frequencies in the corpus showed a wide range because the variance was 64,006.95, while the frequency range extended from 15.73 to 2,345.57 (range = 2,329.84). The word frequency distribution in the corpus showed that few words appeared frequently, but most words appeared rarely.

The Movie Corpus showed 119.81 as its mean expected frequency for 761 analyzed terms, while the standard deviation was 208.05. The word occurrences showed significant variation because of their wide distribution. The frequency distribution showed that words appeared with a median of 48.33, but most words appeared at a frequency of 14.71. The variance was 43,229.01, while the frequency range extended from 12.93 to 1,927.62 (range = 1,914.69). The word frequency distribution showed broad variation because few dominant terms appeared frequently, yet most words appeared rarely throughout the corpus.

The analyzed terms in the Corpus of American Soap Operas showed a mean expected frequency of 157.29 with a standard deviation of 273.15, which indicated substantial word occurrence differences. The frequency distribution showed that words appeared with some regularity because the median frequency was 63.45, and the mode was 19.31. The variance was 74,512.89, while the frequency range extended from 16.97 to 2,530.76 (range = 2,513.79). The majority of words in the corpus appeared rarely because a few dominant terms made up most of the

content. These statistics demonstrated that word frequencies were spread widely throughout the corpus because few dominant terms appeared frequently, but most words appeared rarely.

The TV Corpus showed a mean expected frequency of 157.95 for its analyzed terms, while the standard deviation was 274.29, which indicated wide variation in word appearances. The median frequency was 63.72, while the mode frequency stood at 19.39, which indicated that few words appeared frequently, but most words appeared rarely. The variance was 75,138.08, while the frequency range extended from 17.04 to 2,541.35 (range = 2,524.31). The statistical data showed that words in the corpus appeared at different rates because few dominant terms appeared frequently, but most words appeared rarely.

The 761 analyzed terms in the Wikipedia Corpus showed a mean expected frequency of 128.42, with a standard deviation of 223.02, which indicated large differences in word appearance rates. The median frequency was 51.81, and the mode frequency was 15.76, which indicated that only a few words appeared frequently, but most words appeared rarely. The variance was 49,670.53, while the frequency range extended from 13.86 to 2,066.26 (range = 2,052.40). These statistical data showed that few dominant terms appeared frequently in the corpus, while most words appeared rarely throughout the corpus.

The analyzed terms in the multidisciplinary academic corpus (Kitjaroenpaiboon et al., 2021b) showed an average expected frequency of 24.23, while their standard deviation was 42.08. The distribution of word occurrences showed significant variation because the median frequency was 9.78, and the mode frequency was 2.97. The variance was 1,768.60, while the frequency range extended from 2.61 to 389.90 (range = 387.29). The word frequency distribution showed strong skewness because few terms appeared frequently, but most terms appeared infrequently throughout the corpus.

The SDUIRC (Kitjaroenpaiboon et al., 2024) showed a mean expected frequency of 543.18 with a standard deviation of 904.77, which indicated that word occurrences varied greatly. The median frequency was 236.50, while the mode reached 63.80, which indicated that few terms appeared frequently, but most terms appeared rarely. The variance was 817,534.42, while the frequency range extended from 56.08 to 8,362.67 (range = 8,306.59). The word frequency distribution in the corpus showed strong skewness because few terms appeared frequently, but most terms appeared rarely.

The 761 vocabulary terms in this research study achieved an average log-likelihood (LL) score of 5,741.10 when compared to nine other reference corpora. The LL values showed wide variation because the standard deviation was 10,178.24. The LL values reached their highest point at 2,284.30, while the most frequent value was 822.93. The variance was 103,460,685.76, while the values extended from 349.78 to 89,878.39 (range = 89,528.61). The 761 terms showed high statistical significance because all values exceeded 15.13, which proved their difference from the nine other reference corpora at $p < 0.0001$ (Dunning, 1993; Rayson et al., 2004; Rayson & Garside, 2000; Rayson, 2002, 2008).

The 761 terms studied in this research produced an average Bayes factor (BF) of 5,584.78 when compared to nine other reference corpora while showing a standard deviation of 10,178.24. The statistical analysis showed that terms contributed differently to the results because the median BF was 2,127.99 and the most frequent value was 666.61. The statistical data showed a variance of 103,460,685.76, while the values extended from 193.47 to 89,722.08 (range = 89,528.61). The statistical results showed that all 761 terms produced values above 10, which proved their essential role in differentiating the SDUIBPC from nine other reference corpora.

The 761 terms studied in this research showed an average effect size for log-likelihood (ELL) of 0.00005, while their standard deviation was 0.00006. The

ELL values showed a median of 0.00003 and a mode of 0.00002, which indicated that several terms produced higher effect sizes, but most terms had minimal impact. The data showed no variation because the variance was zero, while the values spanned from 0.00001 to 0.00053 (range = 0.00052). The ELL values from all 761 terms exceeded zero, which proved their statistical distinction from nine other reference corpora, thus establishing their essential role in the SDUIBPC.

The four international baking experts evaluated the importance of 761 terms, which received an average score of 3.93. The experts demonstrated high agreement through their evaluations because the standard deviation was 0.11. The experts assigned a median score of 4.00 and a mode of 4.00, while the variance remained at 0.01. The experts scored between 3.67 and 4.00 on the scale (range = 0.33). The expert panel determined these 761 terms as important because their lowest average score reached 3.26.

The statistical tests and expert evaluations produced results that supported the same conclusion. The 761 terms in SDUIBPC appeared frequently while demonstrating clear importance for international baking language. The high LL and BF values demonstrated that these words appeared with high precision compared to the nine reference corpora which indicated their unique position in this bakery genre. The small ELL in this large dataset followed the statistical patterns that emerged from the analysis. The expert panel showed complete agreement about these terms because they rated them 3.93 on average with only 0.11 standard deviation. The research demonstrated that the 761 identified words represented specialized vocabulary which demonstrated both statistical strength and practical value for bakery language.

The research team identified 761 common words in the SDUIBPC before using established frameworks from Anward (2000), Bisang (2011), Evans (2000), Heine and Kuteva (2012), and Sim and Haspelmath (2012) to classify their functions. The Multidimensional Analysis Tagger (MAT) 1.3.3 (Nini, 2021)

generated classifications which two international baking experts and two native English-speaking syntax faculty members verified for accuracy. The high level of inter-rater agreement confirmed the accuracy and consistency of the functional assignments. These findings support earlier reports that the MAT 1.3.3 achieves accuracy rates above 98% in word function classification (Kitjaroenpaiboon et al., 2024). The functional analysis revealed that the 761 words fall into four main categories: nouns, verbs, adjectives, and adverbs. Of these, 103 words were identified as verbs. In this research, these verbs were grouped by the four experts according to their communicative roles in bakery discourse. This was done for ease of presentation rather than as part of the original analytical procedure.

1. **Preparation of Ingredients** (14 verbs): absorb, adjust, apply, crack, crumble, cube, cut, dice, grind, half, peel, pinch, slice, soak
2. **Incorporation/Mixing** (14 verbs): add, blend, combine, dissolve, incorporate, knead, mix, sift, stir, whisk, whip, beat, overmix, pour
3. **Shaping, Portioning, and Handling** (24 verbs): assemble, divide, drop, fold, form, lay, mark, pipe, place, press, press down, pull, roll, rub, sandwich, shape, spoon, spread, stretch, swirl, transfer, trim, twist, wrap
4. **Core Transformation Processes** (17 verbs): bake, caramelize, chill, evaporate, extract, ferment, freeze, melt, preheat, proof, refrigerate, rise, set, thaw, toast, turn, invert
5. **Decoration, Presentation, and General Instruction** (34 verbs): brush, coat, decorate, drizzle, dust, finish, frost, garnish, ice, scrape, scrape down, serve, spray, sprinkle, check, measure, remove, reserve, use, yield, prepare, put, put in, fill, flip, cover, dip, grease, level, line, separate, soften, thicken, toss

This categorization highlights how verbs in the SDUIBPC were distributed according to the stages of baking, from ingredient preparation, mixing, and shaping, to core transformation, decoration, presentation, and general instructional actions.

5. Discussion

The SDUIBPC analysis used AntConc 4.3.1 (Anthony, 2024) to perform word and n-gram functions with frequency and range measurements. This analysis received additional support from statistical methods which included log-likelihood, Bayes factor, effect size for log-likelihood, and expert opinions in the field. The corpus contained 761 lexical items which appeared frequently throughout the text.

The vocabulary differences between this research and the nine other reference corpora arose from the unique characteristics of text types within each corpus. The nine reference corpora containing spoken and written languages served as general and specialized language corpora. The results of vocabulary analysis depend on the specific texts that researchers collect for the studies (Gries, 2009). The vocabulary patterns in different text types show distinct characteristics because they contain specialized terminology from particular fields which appear in specific genres (Ranney, 2012).

The vocabulary related to baked goods showed a statistically significant difference when researchers compared the 761 words from this study to the nine other reference corpora. The log-likelihood (LL) analysis showed that these differences stemmed from systematic patterns which existed between corpora (Dunning, 1993; Rayson et al., 2004). The Bayes factor (BF) analysis measured the evidence strength for these differences (Kass & Raftery, 1995) and the effect size for log-likelihood (ELL) measured the size and practical value of the results (Paquot, 2007). The language patterns in each corpus arose from their distinct communication settings and functional needs and linguistic features (Kitjaroenpaiboon et al., 2021a, 2021b, 2024). The SDUIRC focuses on international food recipes through procedural communication which includes explanations and preparation instructions for international dishes. The COCA and the GloWbE corpora contain diverse content from various contexts (Davies, 2008, 2013). The Wikipedia Corpus functions as an information source which provides encyclopedic content (Kilgarriff et al., 2010). The Movie Corpus and the American

Soap Opera Corpus contain mostly conversational and narrative content, which shows baking-related words at lower frequencies (Davies, 2008). The multidisciplinary corpus (Kitjaroenpaiboon et al., 2021b) contains academic writing which includes baking terminology only in particular situations. The LL, BF, and ELL results demonstrate that the SDUIBPC contains distinct vocabulary which arise from its specialized baking discourse that involves procedural and instructional content.

The SDUIBPC discovered 761 common words, a total which falls between the 112 terms from Nordin et al. (2013), the 261 terms from Rungrueang et al. (2022), and the 1,165 terms from Kitjaroenpaiboon et al. (2024). This is because the different corpus sizes, domain-specific content, and research methods used in each study created the observed differences in their results. The SDUIBPC analyzed 5.48 million words from 100 international bakery books to create a larger and more specialized vocabulary database than the two previous studies. The study used multiple statistical criteria (i.e., frequency, range, LL, BF, and ELL) together with expert evaluation to achieve a complete identification of important lexical elements (Xodabande et al., 2023). The SDUIRC contains more than 7 million words, which explains why this study identified fewer lexical items than Kitjaroenpaiboon et al. (2024) did.

The SDUIBPC contained 761 lexical items which underwent functional analysis to show that verbs constituted a total of 103 items, ranking as the second most common category after nouns. 11.43% of verbs in the corpus function as essential elements for procedural knowledge representation because they describe the sequential operations which organize baking preparation. The research by Kitjaroenpaiboon et al. (2024) supports this finding because verbs appear as the second most common word type in culinary texts, which contain step-by-step instructions. The analyzed verbs consisted of fundamental baking operations including preparation steps (e.g., mix, measure, knead), cooking methods (e.g., bake, toast), and final presentation elements (e.g., decorate,

garnish, serve). The high occurrence of these verbs demonstrates how bakery discourse depends on specific instructional language, which often appears in imperative form to achieve precise recipe execution (Gorlach, 2004).

The SDUIBPC contains many verbs which also appear in Michael West's General Service List (GSL) from 1953 based on a 5-million-word corpus of written texts. The GSL contains all essential vocabulary for English language learners because it includes fundamental terms that describe basic actions and processes (West, 1953). The GSL contains the same verbs as the SDUIBPC, which include *add*, *combine*, *cut*, *mix* and *pour*. The shared vocabulary between baking discourse and general English usage indicates a strong connection between these two language domains. Gerhardt (2013) explains that eating and speaking represent common human behaviors which people from different cultures and communities share. The words used to describe basic actions including baking terminology stem from human communication needs in everyday life. The appearance of shared verbs between general English and baking-specific contexts demonstrates how language creates connections between different communication areas.

The research found that six verbs from this study (*adjust*, *assemble*, *extract*, *incorporate*, *remove*, and *transfer*) match the Academic Word List (AWL) which Coxhead (2000) established. The AWL contains vocabulary that appears in academic texts but excludes common words from West's General Service List (GSL). The six verbs appear in both the SDUIBPC and the AWL because they function as general procedural verbs which represent fundamental physical and cognitive operations. The verbs demonstrate flexibility because they can explain laboratory work and engineering tasks and cooking methods, which explain their appearance in academic and specialized fields (Biber et al., 1999; Coxhead, 2000). The SDUIBPC contains baking-specific terms including *whisk*, *knead*, *proof*, and *preheat*, which differ from the general vocabulary found in the AWL. The two lists serve different purposes because the AWL helps readers understand academic

content, but the SDUIBPC focuses on bakery-specific terminology needed for professional practice and procedure description.

The SDUIBPC contains shared verbs with previous research studies that focused on specialized vocabulary. The research of Nordin et al. (2013) about food writing teaching vocabulary in Malaysia identified five common verbs with this study. The research by Rungrueang et al. (2022) about food service vocabulary shared 20 terms with this study. The research by Kitjaroenpaiboon et al. (2024) about international food recipe English vocabulary revealed 86 common terms with this study. The overlap verbs across these studies likely arose from their general nature of describing common food preparation steps and instructional commands found throughout different food-related situations.

The 14 bakery-specific verbs from this research study (*absorb, apply, check, frost, invert, level, overmix, press down, proof, put in, scrape down, stretch, swirl,* and *twist*) appeared in no previous studies. The specialized nature of these verbs makes them specific to bakery recipes because they describe unique procedures and techniques found in this particular field. The absence of these terms in previous studies resulted from variations between corpus sizes and source materials, and research focuses which included student food writing, general food service, and wide-ranging recipe collections.

The SDUIBPC analysis showed that the top 10 most common verbs in the corpus included *add, mix, place, bake, use, set, stir, remove, cut* and *extract* which appeared between 7,354 and 17,915 times throughout the texts while maintaining a wide distribution between 87 and 100 texts.

The most common verb in the texts was *add*, which appeared 17,915 times and in all 100 texts. The verb *mix* appeared 14,419 times in the texts because it refers to the process of combining ingredients, which is a crucial step in cooking and appears frequently in instructional writing (Culpeper & Kyto, 2010). The verb

place appeared 13,389 times in the texts because it helps writers specify exact locations (e.g., “place dough on tray”), which matches the technical nature of instructional writing (Hyland, 2021). The domain-specific verb *bake* appeared 12,775 times in the texts because it represents the central cooking process which distinguishes baking texts from other instructional materials (Tribble, 2017).

The verbs *use* and *set* appeared 11,923 and 9,099 times respectively in the texts because they represent the combination of common and specialized terms, which characterizes academic writing (Xiao, 2011). Mid-ranked verbs such as *stir* (frequency = 8,735, range = 98), *remove* (frequency = 8,469, range = 98), and *cut* (frequency = 7,964, range = 96) highlight fine-grained ingredient handling central to recipe cohesion (Gotz & Mukherjee, 2019). The term *extract* appeared 7,354 times throughout the texts with a range of 87. This word indicates specific methods for extracting flavors and essences, the terms for which serve as important lexical markers in baking communication (Biber & Conrad, 2009).

6. Conclusion

The research team of this study generated the SDUIBPC to study its common vocabulary, which focused on verbs. The analysis of 761 frequent words revealed 103 verbs which function as essential guidance for baking procedures including ingredient preparation, mixing, shaping, baking, and decorating. The research results may help teachers to create effective language lessons which allow students to study real-world specialized language patterns from authentic texts. This study also demonstrates how corpus-based methods can be used to create practical applications for teaching domain-specific language and developing recipe content and instructional materials.

7. About the Authors

Associate Professor Dr. Woravit Kitjaroenpaiboon is a Suan Dusit University linguist specializing in corpus linguistics, translation, with extensive editorial experience and numerous scholarly publications.

E-ISSN: 2287-0024

Professor Dr. Sirote Pholpuntin is a scholar in social sciences with extensive leadership experience, currently serving as Administrative Advisor to the President of Suan Dusit University.

Associate Professor Dr. Sukhum Chaleysub is an experienced educator with advanced training in education and information technology, currently serving as Administrative Advisor to the President of Suan Dusit University.

Assistant Professor Samniang Fahkrajang is a language education specialist with extensive teaching experience, currently serving on the Elementary Education Program Committee at Suan Dusit University.

Assistant Professor Prissana Fongsrun is a linguistic scholar with advanced training in linguistic studies, currently serving on the Language and Culture B.A. Program Committee at Suan Dusit University.

Dr. Niphatchanok Najpinij is a gastronomy specialist with interdisciplinary training in communication, hospitality, and culture studies, currently serving as a lecturer and culinary arts expert at Suan Dusit University.

Dr. Chanchana Siripanwattana is a food science educator trained in food technology and culinary arts, now leading curriculum development and teaching food science at Suan Dusit University.

8. Acknowledgement

We gratefully acknowledge Suan Dusit University for supporting this research and the development of the SDUIBPC. We thank our colleagues, collaborators, and experts in international pastry, bakery, as well as our research assistants, for their invaluable guidance and assistance. All remaining errors are ours alone.

9. References

Ambele, E. A. (2022). Thai English? Non-Thai English lecturers' perceptions of Thai English and world Englishers. *LEARN Journal: Language Education and Acquisition Research Network*, *15*(2), 724–750. https://so04.tci-thaijo.org/index.php/LEARN/article/view/259949

Anthony, L. (2024). *AntConc* (Version 4.3.1) [Computer software]. Waseda University. https://www.laurenceanthony.net/software/antconc/

Anward, J. (2000). A dynamic model of part-of-speech differentiation. In P. Vogel & B. Comrie (Ed.), *Approaches to the Typology of Word Classes* (pp. 3–46). De Gruyter Mouton. https://doi.org/10.1515/9783110806120.3

Ari, O. (2006). Review of three software programs designed to identify lexical bundles. *Language Learning & Technology*, *10*(1), 30–37. https://doi.org/10.64152/10125/44044

Azanedo, L., Garcia-Garcia, G., Stone, J., & Rahimifard, S. (2020). An overview of current challenges in new food product development. *Sustainability*, *12*(8), 1–4. https://doi.org/10.3390/su12083364

Babazade, Y. (2024). The impact of digital tools on vocabulary development in second language learning. *Journal of Azerbaijan Language and Education Studies*, *1*(1), 35–41. https://doi.org/10.69760/jales.2024.00103

Baoya, Z. (2015). *Moves and inter-move linguistic variation in education research articles* [Unpublished doctoral dissertation]. Suranaree University of Technology.

Barlow, M. (2000). *MonoConc Pro* (Version 2.0) [Computer software]. Athelstan. http://www.athel.com

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.

Biber, D., Conrad, S., & Leech, G. (2002). *Longman student grammar of spoken and written English*. Longman.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Pearson Education.

Bisang, W. (2011). Word classes. In J. Song (Ed.), *The Oxford handbook of linguistic typology* (pp. 280–302). Oxford University Press.

Can, S., Karabacak, E., & Qin, J. (2016). Structure of moves in research article abstracts in applied linguistics. *Publications*, *4*(3), 1–16. https://doi.org/10.3390/publications4030023

Cargill, M., & O'Connor, P. (2009). *Writing scientific research articles: Strategies and steps*. Wiley–Blackwell.

Chatterjee, U., Kumar, V., & Madalli, D. (2016). Formalizing food ingredients for data analysis and knowledge organization. *Journal of Scientometrics and Information Management*, *10*(2), 289–309. https://doi.org/10.1080/09737766.2016.1213970

Cheevanon, N. (2022). *Factors influencing consumers to select bakery goods in Bangkok* [Master's thesis, Mahidol University]. Mahidol University Institutional Repository. https://archive.cm.mahidol.ac.th/bitstream/123456789/4634/1/TP%20EM.007%202022.pdf

Conrad, S. (1999). The importance of corpus-based research for language teachers. *System*, *27*(1), 1–18. https://doi.org/10.1016/S0346-251X(98)00046-3

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Culpeper, J., & Kyto, M. (2010). *Early modern English dialogues: Spoken interaction as writing*. Cambridge University Press.

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The academic spoken word list. *Language Learning*, *67*(4), 959–997. https://doi.org/10.1111/lang.12253

Davies, M. (2008, May 25). *The corpus of contemporary American English.* English-Corpora. https://www.english-corpora.org/coca/

Davies, M. (2013, June 24). *Corpus of global web-based English (GloWbe)*. English-Corpora. https://www.english-corpora.org/glowbe/

Dessev, T., Lalanne, V., Keramat, J., & Jury, V. (2020). Influence of baking conditions on bread characteristics and acrylamide concentration. *Journal*

of Food Science and Nutrition Research, *3*(4), 291–310. https://doi.org/10.26502/jfsnr.2642-11000056

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61–74. https://aclanthology.org/J93-1003/

Endoo, P. (2017). Corpus linguistics: Principles and applications. *Romphruek Journal*, *35*(1), 164–174. https://romphruekj.krirk.ac.th/wp-content/uploads/sites/2/2020/09/RomphruekJournal35-1.pdf

Evans, N. (2000). Word classes in the world's languages. In G. Booij, C. Lehmann, & J. Mugdan (Eds.), *Morphology: An international handbook on inflection and word-formation* (pp. 708–731). Walter de Gruyter.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, *35*(3), 305–327. https://doi.org/10.1093/applin/amt015

Gerhardt, C. (2013). *Culinary linguistics*. John Benjamins.

Getkham, K. (2014). Politeness strategies in Thai graduate research paper discussions: Implications for second/foreign language academic writing. *English Language Teaching*, *7*(11), 159–167. https://doi.org/10.5539/elt.v7n11p159

Gilner, L., & Morales, F. (2010). Corpus-based frequency profiling: Migration to a word list based on the British National Corpus. *The Buckingham Journal of Language and Linguistics*, *1*, 1–17. https://www.researchgate.net/publication/335802775_Corpus-Based_Frequency_Profiling_Migration_To_A_Word_List_Based_On_The_British_National_Corpus

Gorlach, M. (2004). *Text types and the history of English*. Walter de Gruyter.

Götz, S., & Mukherjee, J. (2019). *Learner corpora and English language teaching*. John Benjamins.

Gries, S. T. (2009). What is corpus linguistics? *Language and Linguistics Compass*, *3*(5), 1225–1241. https://doi.org/10.1111/j.1749-818X.2009.00149.x

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (4th ed.). Advanced Analytics.

Halliday, M. A. K. (1992). Language as system and language as instance: The corpus as a theoretical construct. In J. Svartvik (Ed.), *Directions in corpus linguistics: Proceedings of Nobel Symposium 82*, *Stockholm*, *4-8 August 1991* (pp. 61–78). De Gruyter Mouton. https://doi.org/10.1515/9783110867275.61

Heine, B., & Kuteva, T. (2012). *The evolution of grammar: Tense, aspect, and modality in the languages of the world*. Oxford University Press.

Hyland, K. (2021). *Teaching and researching writing* (3rd ed.). Routledge.

Jones, R., & Tschirner, E. (2015). *A frequency dictionary of German: Core vocabulary for learners*. Routledge.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kilgarriff, A., Reddy, S., Pomikálek, J., & Avinesh, P. V. S. (2010). A corpus factory for many languages. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, *Malta*, *904–910.* https://aclanthology.org/L10-1044/

Kitjaroenpaiboon, W., & Getkham, K. (2015). An analysis of interactional metadiscourse devices in communication arts research articles. *International Journal of Management and Applied Science*, *1*(9), 125–131. https://www.iraj.in/journal/journal_file/journal_pdf/14-196-1446198249125-131.pdf

Kitjaroenpaiboon, W., & Getkham, K. (2016). Stylistic patterns in language teaching research articles: A multidimensional analysis. *PASAA Journal*, *52*, 169–208. https://doi.org/10.58837/CHULA.PASAA.52.1.7

Kitjaroenpaiboon, W., & Getkham, K. (2017). Patterns of linguistic feature and their communicative functions in nursing research articles. *International Journal of Management and Applied Science (IJMAS)*, *3*(3), 98–103.

Kitjaroenpaiboon, W., Khamsakul, B., Kesprathum, S., Fahkrajang, S., & Fongsrun, P. (2021a). Rhetorical move and multidimensional analyses of applied linguistics research abstracts. *Journal of Language and Culture*, *40* (2), 137–165. https://so03.tcithaijo.org/index.php/JLC/article/view/257578

Kitjaroenpaiboon, W., Pholpuntin, S., Fahkrajang, S., Fongsrun, P., Najpinij, N., & Sriarun, J. (2024). *An analysis of international food processing lexis and formulaic languages: An application of a corpus linguistic-based approach* [Unpublished manuscript]. Suan Dusit University.

Kitjaroenpaiboon, W., Wongwiseskul, S., Puksa, T., & Khamsakul, B. (2021b). A multidisciplinary corpus-based comparative analysis: Lexical bundles in language teaching, health sciences, and business management research articles. *NIDA Journal of Language and Communication*, *26*(40), 41–68.

Martínez- Monzó, J., García-Segovia, P., & Albors-Garrigos, J. (2013). Trends and innovations in bread, bakery, and pastry. *Journal of Culinary Science & Technology*, *11*(1), 56–65. https://doi.org/10.1080/15428052.2012.728980

Moussu, L., & Llurda, E. (2008). Non-native English-speaking English language teachers: History and research. *Language Teaching*, *41*(3), 315–348. https://doi.org/10.1017/S0261444808005028

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins.

Nation, I. S. P., & Kyongho, H. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System*, *23*(1), 35–41. https://doi.org/10.1016/0346-251X(94)00050-G

Neuendorf, K. A. (2002). *The content analysis guidebook*. SAGE Publications.

Nguyen, C., & Miller, J. (2020). A corpus-based list of commonly used English medical morphemes for students learning English for specific purposes. *English for Specific Purposes*, *58*, 102–121. https://doi.org/10.1016/j.esp.2020.01.004

Nini, A. (2021). *Multidimensional Analysis Tagger* (Version 1.3.3) [Computer software]. https://sites.google.com/site/multidimensionaltagger/versions

Nordin, M. N. R., Stapa, S. H., & Darus, S. (2013). Developing a specialized vocabulary word list in a composition culinary course through lecture notes. *Advances in Language and Literary Studies*, *4*(1), 78–88. https://doi.org/10.7575/aiac.alls.v.4n.1p.78

Ostonova, S., & Xikmatovna, X. (2020). English language in a historical perspective and forming of vocabulary. *ACADEMICIA: An International Multidisciplinary Research Journal*, *10*(5), 1488–1492. https://doi.org/10.5958/2249-7137.2020.00483.8

Palmer, H. E. (1938). *A grammar of English words*. Longmans, Green and Co.

Paquot, M. (2007). Towards a productively-oriented academic word list. In P. Bouillon, B. Daille, K. Morita, & E. S. M. Tseng (Eds.), *Proceedings of the 4th International Workshop on Computational Terminology* (pp. 49–56). Association for Computational Linguistics.

Puangmali, S. (1976). A study of engineering English vocabulary. *RELC Journal*, *7*(1), 40–52. https://doi.org/10.1177/003368827600700103

Qurboniyozovna, B. (2025). Corpus-based analysis of academic English vocabulary in student writings. *Educational Insight*, *1*(9), 102–120. https://brightmindpublishing.com/index.php/EI/article/view/1378/1406

Ranney, S. (2012). Defining and teaching academic language: Developments in K-12 ESL. *Language and Linguistics Compass*, *6*(9), 560–574. https://doi.org/10.1002/lnc3.354

Rayson, P. (2002). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison* [Unpublished doctoral dissertation]. Lancaster University.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, *13*(4), 519–549. https://doi.org/10.1075/ijcl.13.4.06ray

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora* (pp. 1–6). Association for Computational Linguistics.

Rayson, P., Berridge, D., & Francis, B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. In *Proceedings of the 7th International Conference on Statistical Analysis of Textual Data (JADT 2004)* (pp. 926–936). Presses Universitaires de Louvain.

Riquelme, N., Robert, P., & Arancibia, C. (2022). Understanding older people's perceptions about desserts using word association and sorting task methodologies. *Food Quality and Preference*, *96*, 1–8. https://doi.org/10.1016/j.foodqual.2021.104423

Rokach, A. (2020). Belonging, togetherness and food rituals. *Open Journal of Depression*, *9*(4), 77–85. https://doi.org/10.4236/ojd.2020.94007

Rungrueang, T., Boonprasert, P., Poempongsajaroen, S., & Laosrirattanachai, P. (2022). Corpus-based approach to generate a word list for food service. *THAITESOL Journal*, *35*(1), 57–76. https://so05.tci-thaijo.org/index.php/thaitesoljournal/article/view/258591

Scott, M. (2024). *WordSmith Tools* (Version 9) [Computer software]. Lexical Analysis Software.

Sicanore, J. M., Connell, K. J., Olthoff, A. J., Friedman, M. H., & Geght, M. R. (1999). A method for measuring interrater agreement on checklists. *Evaluation & the Health Professions*, *22*(2), 221–234. https://doi.org/10.1177/01632789922034284

Sim, S., & Haspelmath, M. (2012). *Understanding morphology* (2nd ed.). Oxford University Press.

Sinclair, J. (2014). Corpus evidence in language description. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 27–39). Routledge. https://doi.org/10.4324/9781315842677-3

Smanalieva, J. (2025). Innovations in food production. *Bulletin of the Kyrgyz National Agrarian University*, *23*(2), 73–83. https://doi.org/10.63621/bknau./2.2025.73

Sonia, S., Sadozai, K., Khan, N., & Jan, A. (2019). Assessing the impact of climate change on wheat productivity in Khyber Pakhtunkhwa, Pakistan. *Sarhad*

Journal of Agriculture, *35*(1), 284–292. https://doi.org/10.17582/journal.sja/2019/35.1.284.292

Sutton, D. (2017). Cooking Skills, The senses and memory: The fate of practical knowledge. In E. Edwards, C. Gosden & R. Phillips. (Eds), *Food and culture* (pp. 75–97). Routledge. https://doi.org/10.4324/9781315680347-7

Tribble, C. (2017). *Writing academic English: Corpus perspectives*. Routledge.

Wang, J., Liang, S., & Ge, G. (2008). Establishment of a medical academic word list. *English for Specific Purposes*, *27*(4), 442–458. https://doi.org/10.1016/j.esp.2008.05.003

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes*, *28*(3), 170–182. https://doi.org/10.1016/j.esp.2009.04.001

Wesser, G. (2021). Thuringian festive cakes: Women's labour of love and the taste of Heimat. In F. Edwards, R. Gerritsen, & G. Wesser (Eds.), *Food, senses and the city* (pp. 108–121). Routledge.

West, M. (1953). *A general service list of English words*. Longman.

Xiao, Y. (2011). [Review of the book *Academic discourse: English in a global context*]. *Journal of English for Academic Purposes*, *10*(3), 198–199. https://doi.org/10.1016/j.jeap.2010.02.009

Xodabande, I., Atai, M. R., & Thompson, P. (2023). Developing and validating a mid-frequency word list for chemistry: A corpus-based approach using big data. *Asian-Pacific Journal of Second and Foreign Language Education*, *8*, 1–21. https://doi.org/10.1186/s40862-023-00205-5

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication*, *3*(2), 215–229.

Zanettin, F. (2014). *Corpora in translation*. Palgrave Macmillan. https://link.springer.com/chapter/10.1057/9781137025487_10

Żurek, J., & Rudy, M. (2024). Impact of the COVID-19 pandemic on changes in consumer purchasing behavior in the food market with a focus on meat and meat products—A comprehensive literature review. *Foods*, *13*(6),1–22. https://doi.org/10.3390/foods13060933