Journal of Multidisciplinary in Social Sciences (September – December 2026), 22(3), e280463
https://doi.org/10.65763/jmss.2026.e280463

Analysis of International Food Processing Lexis and Formulaic Languages: A Corpus Linguistic-Based Approach

Woravit Kitjaroenpaiboon[a], Sirote Pholpuntin[b] , Samniang Fahkrajang[a,] Niphatchanok Najpinij[b], Prissana Fongsrun[a] & Jareuk Sriaroon[c]

[a] Faculty of Humanities and Social Sciences, Suan Dusit University, Bangkok, 10300 Thailand
[b] Graduate School, Suan Dusit University, Bangkok, 10300 Thailand
[c] School of Culinary Arts, Suan Dusit University, Bangkok, 10700 Thailand

Abstract

Gastronomy plays an important cultural role and often attracts people from diverse cultural backgrounds. However, learners still face difficulties using English effectively when discussing culinary topics. Therefore, this study investigates the English words frequently used in international culinary books by compiling and analyzing a specialized corpus together with the AntConc 4.3.1 of Anthony (2024). Log Likelihood, Bayesian Information Criterion based Bayes Factor, Effect Size for Log Likelihood, and expert evaluation were together employed to identify statistically significant lexical items. The analysis revealed 1,165 English words that frequently occur in international recipe books. The cluster validation demonstrated high reliability (93.61%), indicating that the identified items consistently co-occur in the corpus. This represents the core vocabulary of the international recipe corpus. The dependent-sample *t-test* results indicated significant differences between these 1,165 items and those in eight reference corpora ($p < .001$). These findings confirm their domain-specific status. The identified 1,165 vocabulary was classified into six functional categories: 1) terms associated with international culinary procedural aspects, 2) terms linked to international culinary ingredients, 3) terms related to international culinary apparatus, 4) terms delineating the attributes of international food and culinary apparatus, 5) terms portraying the characteristics of international culinary processes, and 6) other pertinent terms related to international food recipes. Also, contextual analysis provided Thai translation of the 1,165 frequent lexical items and examined 1,192 verb-collocations. The collocation analysis shows that many vocabulary items regularly co-occur with specific procedural verbs. This study offers a lexical resource that can support language learning and communication in the international culinary domain.

Keywords: Corpus linguistics, International recipes, Qualitative descriptors, Adverbial modifiers

*** Corresponding Author**
e-mail: niphatchanok_naj@dusit.ac.th
ORCID iD: https://orcid.org/009-002-5265-6696

Introduction

Culinary culture is widely celebrated across the world. Many travelers actively explore diverse cuisines as part of their cultural experiences (Risteski & Korunovski, 2016). English functions as a global medium of communication in tourism, business, science, and education (Sudarmo, 2021). However, in countries where English is not widely used in everyday communication, many people continue to experience difficulties using specialized English vocabulary for communication effectively (Moussu & Llurda, 2008). Although Thai cuisine has internationally achieved strong recognition, limited proficiency in specialized English culinary vocabulary and their formulaic language may constrain the ability of Thai people to communicate culinary knowledge, to participate in international culinary education, and to globally promote Thai food culture. Because vocabulary and formulaic language knowledge constitutes a central component of language proficiency and supports the development of listening, speaking, reading, and writing skills (Ricketts et al., 2007), the systematic identification of domain-specific lexis and formulaic languages are both particularly important in contexts such as Thailand, where English functions neither as the first nor the second language (Kitjaroenpaiboon & Getkham, 2015, 2016, 2017).

Corpus linguistics, which analyzes authentic language data drawn from natural texts, provides a reliable analytical framework for identifying the lexis and formulaic languages frequently occurred in domain-specific

communication (Aroonmanakun, 2011). In culinary contexts, understanding food-processing lexis is essential for accurately interpreting and communicating recipes, procedural manuals, and food materials (Rashid et al., 2022; Suwonsichon, 2019). Formulaic expressions enhances communicative fluency because they are processed and retrieved as prefabricated linguistic units (Conklin & Schmitt, 2012). They also contribute to the standardization, clarity, and efficiency of procedural instructions in professional culinary communication (Wray, 2002). Consequently, mastery of both specialized vocabulary and collocations is fundamental for effective culinary education, professional training, and international knowledge exchange.

Previous corpus-based studies (e.g., Nordin et al., 2013; Rungrueang et al., 2022) have contributed culinary vocabulary lists from food-article-writing instructional materials and food-service corpora. These investigations identify lexical items based on frequency, range, and keyword analyses, providing useful wordlists for specialized instruction. Despite these contributions, existing studies rarely integrate lexical frequency, functional classification, Thai interpretation, and collocations within a single corpus-driven framework. Most existing studies focus primarily on vocabulary compilation and devote comparatively limited attention to the broader lexical-functional organization of culinary discourse, the integration of translational accessibility for Thai users, or the systematic examination of collocations that structure procedural meaning. As culinary discourse continues to evolve through technological innovation, global recipe circulation, and changing food-processing practices, updated corpus-based analyses that integrate lexical frequency, functional classification, Thai interpretation, and collocational patterning are increasingly necessary.

Addressing this gap and expanding knowledge edge, the present study investigates frequent lexical items in an international corpus of English-language recipe books in order to identify the core lexis that structure culinary discourse. The study integrates several complementary analyses, including frequency and dispersion-based lexical identification, cross-corpus comparison, functional lexical classification, Thai interpretation of lexical items, and the examination of verb–collocations that encode frequent procedural expressions. By combining these approaches within a unified corpus-linguistic framework, the study aims to provide a comprehensive description of the lexical structure and communicative functions of international culinary discourse while also developing practical linguistic resources to support culinary education, translation, and professional communication in bilingual contexts.

Objectives

This study aims to

1) identify high-frequency lexical items in the Suan Dusit University International Recipe Corpus;

2) compare these lexical items with reference corpora to determine statistically distinctive features of culinary discourse;

3) classify the lexical items according to their functional categories to construct a functional profile of international recipe language;

4) provide context-based Thai equivalent interpretations to enhance cross-linguistic accessibility; and

5) examine verb-based collocations to identify recurrent formulaic constructions that characterize procedural expressions in culinary instructions.

Through these integrated objectives, the study seeks to provide a comprehensive account of specialized culinary lexis and formulaic language in international recipe discourse.

Conceptual framework

This study drew on established vocabulary-analysis frameworks (Chung & Nation, 2004; Coxhead, 2000; Dang et al., 2022; Laosrirattanachai & Ruangjaroon, 2020) to guide the identification and compilation of lexical items from one hundred international recipe books, while adopting the functional classification approach of Nordin et al. (2013) to categorize food-processing lexis. To enhance cross-linguistic accessibility for Thai users, Thai equivalent meanings were provided for lexical items, with translations derived from corpus-based contextual analysis (Baker; 2011; Newmark, 1998) and validated through expert review and consultation of dictionaries. In addition, the study examined verb–collocations using AntConc 4.3.1 (Anthony, 2024), applying dispersion thresholds and Mutual Information (MI) statistics following Brezina et al. (2015) to identify statistically significant verb-collocations. Interpreting MI values ($MI \geq 3.0$) alongside dispersion measures ($Range \geq 5$) ensures that the resulting collocations reflect both strong lexical association and consistent distribution across the corpus.

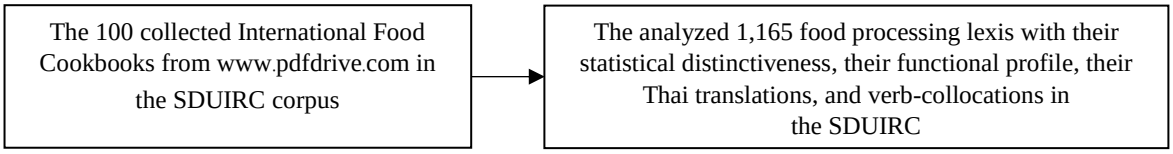


Figure 1 Conceptual framework

Methodology

This study received exemption from ethical approval under certification number SDU-RDI-SHS 2024-034.

Corpus Generation

A specialized language corpus, Suan Dusit University International Recipe Corpus (SDUIRC), was developed for this study. The corpus was compiled from international English-language recipe books selected from the top one hundred most downloaded books available on *www.pdfdrive.com*. The texts were accessed solely for academic research purposes, and no materials were redistributed or reproduced beyond analytical use. Selection was based on download rankings to ensure the inclusion of widely accessed and frequently consulted culinary texts. Duplicate titles were excluded. The resulting corpus contained 7,090,414 words and focused exclusively on savory dishes, with dessert and beverage recipes excluded in order to maintain thematic consistency in the analysis of cooking-process discourse. The use of *www.pdfdrive.com* as the primary source was motivated by the platform's global accessibility and extensive user base, which reflects the reading preferences and practical interests of an international audience. By prioritizing highly downloaded book titles, the corpus captures culinary texts that are widely circulated and likely to influence cross-cultural culinary communication. Although this sampling strategy may introduce certain regional or platform-related biases, it also reflects authentic patterns of digital language exposure and engagement. This perspective is in line with Tenopir et al. (2012). Consequently, the SDUIRC provided a substantial and valid dataset for examining widely disseminated English-language recipe discourse.

Data Processing and Reliability

To ensure the accurate categorization of recipe types in the corpus, four researchers and four culinary experts (two professional chefs and two culinary educators) participated in discussions. Recipe classification followed the ideas proposed by Galindo-Cuspinera & Breslin (2006) and Wang et al. (2019), which divides recipes into two principal categories: savory (including salty and spicy dishes) and sweet (including recipes containing sugar, honey, fruits, chocolate, or confectionery ingredients). After an initial calibration discussion to establish a shared understanding of the classification criteria, each expert independently classified a sample of fifty recipes. Intercoder agreement was calculated using percentage agreement, with a minimum threshold of 80% required for acceptance. Discrepancies were resolved through group discussion, thereby strengthening the reliability of the corpus-based classification procedure. All selected texts were converted into machine-readable plain-text format before the analyses. Non-linguistic elements such as images, page numbers, formatting symbols, and publisher metadata were removed to ensure analytical consistency. The cleaned corpus was segmented into individual recipe texts and standardized through normalization procedures, including the correction of character-encoding inconsistencies and typographical errors introduced during file conversion. Word-token counts and text segmentation were verified before lexical analysis to ensure corpus integrity.

Lexical Extraction

Lexical extraction was conducted using AntConc 4.3.1 (Anthony, 2024). Frequent food-processing lexical items were operationally defined as those occurring at least five times per 100,000 words and appearing in a minimum of five texts. Given the corpus size of 7,090,414 words, qualifying items were therefore required to occur at least 355 times across at least five texts. Lexical identification followed established corpus-linguistic procedures, including frequency and dispersion criteria. Log-likelihood (LL), Bayesian Information Criterion based Bayes Factor (BICBF), and Effect-size Log-likelihood (ELL) were employed to ensure statistical robustness. Items were retained in the final list only when statistical thresholds were satisfied (LL ($p < .05$); BICBF > 10; ELL values exceeding zero, indicating substantive effect strength), thereby ensuring that statistically significant items also demonstrated substantive practical importance. Expert evaluation, with mean ratings ranging from 3.26 to 4.00, was conducted to assess the validity of the extracted items (Chung & Nation, 2004; Coxhead, 2000; Dang et al., 2022; Laosrirattanachai & Ruangjaroon, 2020), and cluster validation was subsequently employed to examine their reliability and internal consistency. The final lexical list was compiled based on converging quantitative and qualitative evidence.

Cross-Corpus Comparison

To determine whether the identified lexical items demonstrate statistically distinctive characteristics of culinary discourse, cross-corpus comparisons were conducted. Paired-samples *t-tests* (or dependent sample *t-test*) were applied to compare the SDUIRC lexical items with eight reference corpora. These analyses established whether the identified items occurred significantly more frequently in the culinary corpus than in general reference corpora, thereby confirming their domain-specific status.

Functional Categorization

Initial grammatical classification of lexical items was conducted based on established traditional grammatical framework of Haspelmath (2012), which recognize eight principal word classes in English: nouns, pronouns, verbs, adjectives, adverbs, prepositions, conjunctions, and interjections. In this study, Multidimensional Analysis Tagger 1.3.3 (Nini, 2019) performed part-of-speech tagging. Then, the tagging results were reviewed through collaborative evaluation involving the research team, two international culinary specialists, and two native-speaker linguists with expertise in syntax. All lexical items were examined within their corpus contexts, and final classifications were confirmed only when agreement was reached between expert judgment and automated tagging results. For greater accessibility (particularly for non-native English users) the study reorganized traditional word classes into meaningful culinary categories following the functional classification approach proposed by Nordin et al. (2013), with minor adaptations. Therefore, lexical items were grouped into six functional categories: 1) terms associated with international culinary procedural aspects, 2) terms linked to international culinary ingredients, 3) terms related to international culinary apparatus, 4) terms delineating the attributes of international food and culinary apparatus, 5) terms portraying the characteristics of international culinary processes, and 6) other pertinent terms related to international food recipes. The categorization process was also conducted by the research team (two culinary experts and two native-speaker linguists) to ensure analytical accuracy, reliability, and disciplinary relevance.

Thai Interpretation

Furthermore, to enhance accessibility for Thai users, Thai equivalent translations were provided for the high-frequency lexical items identified in the corpus. Translation procedures were guided by established translation frameworks (i.e., Baker, 2011; Newmark, 1998), prioritizing meaning equivalence while permitting necessary structural adjustments to preserve contextual meaning. Preliminary interpretations were derived through contextual analysis (i.e., examination of their Thai meanings within the surrounding textual environment) and subsequently evaluated by the research team, culinary specialists, and experts in linguistics and English-Thai translation. Candidate translations were further verified through consultation with authoritative general and specialized culinary dictionaries and additional corpus resources, and final interpretations were accepted only after expert consensus had been achieved.

Collocation Identification

Verb-collocations were examined to determine how frequently occurring verbs combine with surrounding lexical items to form conventionalized instructional expressions in recipes. Verb-collocations were prioritized because they serve as the central organizing elements of instructional discourse, encoding procedural actions and governing the syntactic and semantic structure of recipe instructions. Verb-collocation analysis was conducted using AntConc 4.3.1 (Anthony, 2024), applying a minimum dispersion threshold (range ≥ 5 texts) together with statistical association measures, particularly Mutual Information ($MI \geq 3.0$), to identify significant collocations. In accordance with corpus-linguistic methodological recommendations, MI scores were interpreted alongside dispersion statistics to ensure that identified collocations reflected both strong lexical association and consistent distribution across the corpus (Brezina et al., 2015).

Results

Based on the predefined frequency and dispersion criteria (minimum five occurrences per 100,000 words and presence in at least five texts), the analysis identified 1,165 frequent lexical items in the SDUIRC.

Frequency analysis conducted using AntConc 4.3.1 (Anthony, 2024) showed that the 1,165 lexical items had an average frequency of 2,700.71 occurrences (SD = 4,974.53), with a median frequency of 1,081 and a mode of 356, indicating substantial variation in lexical prominence across the corpus. Frequencies ranged from 355 to 67,556 occurrences. These reflect the coexistence of both moderately frequent and highly dominant lexical items within recipe discourse.

Dispersion analysis further indicated that the 1,165 lexical items appeared in an average of 63.48 texts (SD = 22.12), with a median of 64 and a range from 5 to 100 texts. These suggest that most identified items were widely distributed across multiple recipe sources rather than concentrated in a limited subset of texts.

To evaluate domain specificity, the identified 1.165 lexical items were compared with eight reference corpora, including the Corpus of Contemporary American English (COCA), the Global Web-based English (GLOWBE) Corpus, the TV Corpus, The Movie Corpus, the Corpus of American Soap Operas, the Wikipedia Corpus, the Coronavirus Corpus, and the Multidisciplinary Academic Corpus. Statistical distinctiveness was assessed using LL statistics ($p < .05$), BICBF, and ELL measures, complemented by expert evaluation.

Across the eight reference corpora, the 1,165 identified vocabulary items exhibited lower mean frequencies and high variability. For instance, the mean frequencies were 158.10 in COCA (SD = 612.65), 101.31 in the Coronavirus Corpus (SD = 340.77), 69.52 in GLOWBE (SD = 182.67), 58.88 in the Movie Corpus (SD = 255.38), 65.02 in the Soap Opera Corpus (SD = 410.01), 78.99 in the TV Corpus (SD = 357.60), 30.36 in the Wikipedia Corpus (SD = 99.93), and 10.34 in the Academic Corpus (SD = 37.02). Median and mode values were frequently near zero. These indicate that many items were domain-specific to the recipe corpus. Combined statistics across corpora yielded a mean frequency of 3,273.24 (SD = 5,579.07), with median and mode values of 1,419 and 654, respectively.

Expected frequencies were also calculated across corpora to determine statistical distinctiveness. LL analysis produced a mean value of 6,797.73 ($p < .0001$). It indicates strong cross-corpus differences. BICBF values averaged 6,728.95 (BICBF > 10). It provides decisive evidence of distinctiveness. The mean ELL value (0.00006) further confirmed that the vocabulary items differed systematically from their distributions in the reference corpora.

To strengthen the domain validity of the vocabulary list, expert evaluation was incorporated into the selection process. Four specialists (two professional chefs and two culinary educators) assessed the relevance of lexical items to the culinary domain, following expert-judgment procedures commonly applied in domain-specific vocabulary research (e.g. Chung & Nation, 2004). Each item was rated using a four-point relevance scale: (1) not related, (2) slightly related, (3) mostly related, and (4) highly specific to the culinary field. Interpretation of mean scores followed the interval-width procedure derived from Fisher's formula (Likert, 1932), calculated as:

$$\textit{Interval width} = \frac{\textit{Maximum value} - \textit{Minimum value}}{\textit{Number of scale points}} = \frac{4-1}{4} = 0.75$$

Accordingly, mean ratings were interpreted as follows: 1.00–1.75 (unrelated), 1.76–2.50 (marginally related), 2.51–3.25 (significantly related), and 3.26–4.00 (highly domain-specific). The expert evaluation produced an overall mean score of 3.93, indicating that the selected lexical items were strongly representative of the culinary domain. This procedure provides systematic expert validation and supports the reliability of the compiled vocabulary set.

The researchers further conducted statistical analyses to examine whether the identified words in the SDUIRC formed meaningful frequency-based groups and to explore whether the lexical items exhibited significant differences compared to those in eight reference corpora. First, cluster validation was performed to determine whether the identified lexical items formed internally consistent frequency-based groupings, thereby assessing the structural reliability of the extracted lexical set. Subsequently, a paired-samples *t-test* was employed to compare normalized frequencies across corpora, a statistical procedure used to compare two related groups. A paired-samples *t-test* was considered appropriate because each lexical item constituted a matched observation across corpora, with its normalized frequency in the SDUIRC directly paired with its corresponding frequency in each reference corpus. To facilitate these two analyses, the researcher took a list of frequently occurring 1,165 terms from the SDUIRC and searched for these terms across eight distinct reference corpora. Before conducting the cluster validation and paired *t-test*, the frequency data for the 1,165 terms in the SDUIRC and in the eight reference corpora were normalized to account for differences in text and corpus size and ensure comparability. The normalization process involved converting raw frequency counts to normalized frequencies. This approach allowed for the standardization of frequency data, enabling a more accurate comparison of term usage across the different corpora. The normalized frequency data for the 1,165 words were examined across the eight reference corpora. The study found that:

The SDUIRC: The mean normalized frequency was 38.07, with a standard deviation of 70.17, a median of 15.23, a mode of 5.02, variance of 4,918.95, a range of 950.21, a maximum frequency of 952.78, and a minimum of 2.57. These confirm substantial variability across lexical items.

Corpus of Contemporary American English (COCA): The mean normalized frequency was 1.68, with a standard deviation of 6.51, a median of zero, a mode of zero, variance of 42.33, a range of 102.41, a maximum frequency of 102.41, and a minimum of zero. These suggest that only a small number of words appear occasionally.

Corona Virus Corpus (COVID): The mean normalized frequency was 3.49, with a standard deviation of 11.73, a median of 0.14, a mode of zero, variance of 137.57, a range of 239.42, a maximum frequency of 239.42, and

a minimum of zero. These suggest a highly uneven distribution, where only a small number of terms appear occasionally at higher frequencies.

Global Web-based English Corpus (GLOWBE): The mean normalized frequency was 3.5, with a standard deviation of 9.18, a median of 0.3, a mode of zero, variance of 84.29, a range of 104.54, a maximum frequency of 104.54, and a minimum of zero. These suggest uneven distribution, with only a small number of words appearing more frequently.

The Movie Corpus (Movie): The mean normalized frequency was 3.60, with a standard deviation of 15.63, a median of 0.18, a mode of zero, variance of 243.95, a range of 304.77, a maximum frequency of 304.77, and a minimum of zero. These show considerable dispersion, indicating that only a small number of items appear with higher frequencies while the majority occur infrequently or not at all.

Corpus of American Soap Operas (Soap Operas): The mean normalized frequency was 3.03, with a standard deviation of 19.11, a median of 0.05, a mode of zero, variance of 364.8, a range of 492.98, a maximum frequency of 492.98, and a minimum of zero. These indicate substantial variability, with only a few items appearing more frequently while the majority remain absent or infrequent.

The TV Corpus (TV): The mean normalized frequency was 3.67, with a standard deviation of 16.6, a median of 0.28, a mode of zero, variance of 275.18, a range of 380.88, a maximum frequency of 380.88, and a minimum of zero. These suggest uneven distribution, with only a few lexical items occurring more frequently while the majority remain infrequent.

The Wikipedia Corpus (Wikipedia): The mean normalized frequency was 1.73, with a standard deviation of 5.7, a median of zero, a mode of zero, variance of 32.5, a range of 79.51, a maximum frequency of 79.51, and a minimum of zero. These suggest some variability, with only a small number of items appearing more frequently.

Multidisciplinary Academic Corpus (Academic) (Kitjaroenpaiboon et al., 2021): The mean normalized frequency was 3.13, with a standard deviation of 11.2, a median of zero, a mode of zero, variance of 125.32, a range of 196.32, a maximum frequency of 196.32, and a minimum of zero. These suggest uneven distribution, with only a small number of items appearing more frequently.

The cluster validation showed that 1,090 (93.61%) of the 1,165 lexical items frequently identified in the SDUIRC consistently co-occurred within the corpus. According to the Kappa statistic framework, agreement levels between 81% and 100% indicate high reliability (McHugh, 2012). Therefore, the high level of agreement suggests that the identified vocabulary reliably represents the core lexical profile of international recipe discourse in the corpus.

Paired-samples *t-tests* were conducted to compare the normalized frequencies of the 1,165 lexical items across corpora. Because each lexical item was measured in both the SDUIRC and the eight reference corpus, each word constituted a matched observation across corpora, thereby justifying the use of a dependent-samples design. The results revealed statistically significant differences between the SDUIRC and all eight reference corpora (all $p < .001$). Specifically, comparisons between the SDUIRC and COCA showed a substantial mean difference ($M = 36.39$, $SD = 69.53$), $t(1164) = 17.87$, 95% CI [32.39, 40.39], with a medium-to-large effect size (Cohen's $d = 0.52$). Similar significant differences were observed for COVID ($t(1164) = 16.92$, $d = 0.50$), GLOWBE ($t(1164) = 17.14$, $d = 0.50$), Movie ($t(1164) = 16.80$, $d = 0.49$), Soap Operas ($t(1164) = 16.86$, $d = 0.49$), TV ($t(1164) = 16.75$, $d = 0.49$), Wikipedia ($t(1164) = 17.81$, $d = 0.52$), and Academic ($t(1164) = 17.08$, $d = 0.50$). These consistent medium-to-large effect sizes indicate that the lexical items identified in the SDUIRC are not only statistically distinct but also substantively characteristic of English-language international recipe discourse.

After extracting vocabulary from the SDUIRC, the researchers applied theoretical frameworks from Haspelmath (2012) to analyze the grammatical functions of the identified vocabulary. The classification process was carried out by Multidimensional Analysis Tagger 1.3.3 (Nini, 2019) and the resulting classifications were subsequently reviewed and validated by the researchers, two international culinary experts and two native English professors, with expertise in syntax. The findings indicate that the frequently occurring lexical items in the corpus perform four principal grammatical functions: nouns, verbs, adjectives, and adverbs. Among the identified items, 790 words (67.81%) function as nouns, 188 words (16.13%) as verbs, 140 words (12.01%) as adjectives, and thirty-four words (2.91%) as adverbs. A small number of items display multifunctionality, including nine words (0.77%) functioning as both nouns and verbs and four words (0.34%) functioning as both verbs and adjectives. The results demonstrate a clear predominance of nominal vocabulary within the SDUIRC, reflecting the strong emphasis on ingredients, materials, and culinary entities characteristic of instructional recipe discourse. This predominance of nominal items indicates the high referential density of recipe discourse, in which ingredients, tools, and food entities are prioritized over interpersonal or evaluative language.

After completing the grammatical classification, the researchers determined that the study should produce knowledge accessible to non-native English users. Accordingly, traditional grammatical categories were reorganized into more user-friendly functional groupings while remaining grounded in established grammatical frameworks. The

reclassification followed the culinary lexical taxonomy proposed by Nordin et al. (2013), which organizes food-related vocabulary into functional culinary domains. In this study, nouns were divided into "terms linked to international culinary ingredients" and "terms related to international culinary apparatus"; verbs were grouped as "terms associated with international culinary procedural aspects"; adjectives were treated as "terms delineating the attributes of international food and culinary apparatus"; and adverbs were categorized as "terms portraying the characteristics of international culinary processes". However, lexical items that could not be assigned to any of these categories were classified under "other pertinent terms related to international food recipes". Please note that, due to the space limitations, only distributional frequencies of lexical categories are presented in this article, while the complete list of the identified 1,165 is provided in the full research report of Kitjaroenpaiboon et al. (2025).

The lexical categorization analysis identified 193 terms (e.g. *add, assemble, blanch, caramelize, dash, etc.*) representing terms associated with international culinary procedural aspects, 464 terms (e.g. *all-purpose flour, banana, canola oil, dry mustard, fish, etc.*) classified as terms linked to international culinary ingredients, 92 terms (e.g. *airtight container, baking dish, colander, electric mixer, fork, etc.*) classified as terms related to international culinary apparatus, 115 terms (e.g. *plain, round, salty, tasty, uncooked, etc.*) classified as terms delineating the attributes of international food and culinary apparatus, 34 terms (e.g. *approximately, coarsely, finely, occasionally, usually, etc.*) classified as terms portraying the characteristics of international culinary processes, and the other 281 terms (e.g. *heavy, high pressure, key nutrient, leftover, magnesium, etc.*) classified as other pertinent terms related to international food recipes. The lexical categorization revealed that fourteen food processing lexis (e.g. *cool, dash, drop, empty, pinch, etc.*) appeared in more than one category, reflecting their multifunctional use across different culinary contexts.

Then, Thai interpretation of the frequently occurring vocabulary was conducted using established translation frameworks (e.g., Baker, 2011; Newmark, 1998), which served as the conceptual basis for determining Thai meanings. Each lexical item was examined across all corpus contexts in which it occurred to identify potential contextual variation in meaning. Proposed interpretations were then cross-checked against authoritative general and specialized culinary dictionaries and additional corpus resources. Final meanings were accepted only when consensus was reached among the research team, two culinary experts and two specialists, in linguistics and English–Thai translation. Where lexical items exhibited multiple meanings across different contexts, all context-specific interpretations were retained. Subsequently, the finalized meanings of the 1,165 lexical items were reviewed by four culinary professionals affiliated with Suan Dusit University to confirm clarity and interpretive accuracy. The study identified Thai equivalent translations for all 1,165 lexical items; however, presenting the complete set of translations is beyond the space of this article. Accordingly only selected representative examples are provided here to illustrate the interpretive approach adopted in the analysis. The complete list of the interpretations is provided in the full research report of Kitjaroenpaiboon et al. (2025). For example, the term "*season*" was interpreted both as a verb meaning "pruŋ˧ rot̚˧/" in Thai (to add seasoning). Similarly, the word "*light*" was interpreted as "t͡ɕɯa̯˧ d͡ʑaːŋ˧" in Thai when referring to diluted; weak; thin (in concentration), not concentrated flavor (e.g., *light sauce*), with each meaning retained according to its contextual usage. The word "*ingredient*" was interpreted as "sùan pʰà.sŏm" in Thai.

The study further investigated the formulaic languages or collocations of 193 high-frequency verbs in English international recipe texts by examining their co-occurrence with surrounding words in the SDUIRC. Verb-collocation analysis was conducted using AntConc 4.3.1 (Anthony, 2024), employing a five-word span, a minimum dispersion threshold of five texts (range ≥ 5), and Mutual Information scores of 3.0 or higher (MI ≥ 3.0) to identify statistically significant and consistently distributed associations. The analysis identified 1,192 verb–collocations meeting the established statistical criteria. Owing to space limitations, only a representative selection of the most frequent and functionally salient collocations is presented in this article while the complete verb-based collocation list is available in the full research report of Kitjaroenpaiboon et al. (2025).

For example, the verb "*add*" frequently co-occurs with ingredient nouns and quantity-related phrases, forming conventional procedural expressions in recipe instructions. Typical patterns include quantity-modifying constructions such as *add a little more oil*, *add a little more water*, and *add a little water*, as well as ingredient-listing constructions such as *add all ingredients*, *add all the ingredients*, *add all remaining ingredients*, and *add all other ingredients*. The verb also regularly combines with specific ingredient nouns, producing expressions such as *add bay leaf, add bell peppers, etc.*

Similarly, the verb "*beat*" frequently co-occurs with ingredient nouns and result-oriented process phrases, forming conventional procedural expressions in recipe instructions. Common ingredient-based combinations include *beat egg whites, beat the egg yolks, beat the eggs, beat the butter, beat the cream,* and *beat the mixture,* all of which describe the mechanical mixing or aeration of ingredients during preparation. In addition, the verb regularly appears in process-result constructions such as *beat until smooth,* which specify the desired texture outcome of the mixing procedure.

The verb "*chop*" regularly co-occurs with ingredient nouns, forming conventional preparation-stage instructions in recipe discourse. Frequent combinations include *chop the garlic* and *chop the onion*, both referring to the cutting of ingredients into smaller pieces prior to cooking.

The verb "*cube*" was found to collocate with ingredient nouns in procedural instructions, forming expressions such as *cube the tofu* and *cube fresh ginger*.

The verb "*dice*" frequently collocates with ingredient nouns, forming procedural expressions such as *dice the avocado*, *dice the apple*, *dice the bacon*, *dice the bell pepper*, *dice the tomato*, *dice the chicken*, *dice the onion*, *dice the tofu*, and *dice the vegetables*

The verb "*garnish*" was also found to occur in recurrent collocational patterns, typically followed by the prepositional phrase *with* and ingredient nouns, forming expressions such as *garnish with basil*, *garnish with chopped parsley*, *garnish with cilantro*, *garnish with fresh parsley*, *garnish with green onions*, *garnish with lemon*, *garnish with parsley*, and *garnish with scallions*.

Conclusion and Discussion

The study identified 1,165 frequently occurring lexical items in the SDUIRC. Multiple statistical indicators (i.e. LL, BICBF, ELL, expert evaluations, and cluster analysis) confirmed that these 1,165 lexical items represent statistically significant and domain-specific key vocabulary in international culinary discourse.

A comparison of the 1,165 lexical items identified in the present study with eight reference corpora shows statistically significant differences in vocabulary distribution. These differences can be explained by corpus type and textual composition. The eight comparison corpora include both spoken and written language and primarily represent general-purpose corpora (e.g., COCA, Wikipedia , Movie , GlOWBE, Soap Opera , TV), as well as a small number of specialized corpora such as the Academic and the COVID. Because vocabulary is strongly influenced by genre, communicative purpose, and discourse domain, each corpus naturally produces distinct lexical profiles. The present findings are in line with Biber's (1988) argument that linguistic variation systematically reflects differences in situational and functional contexts across discourse types.

In contrast, the SDUIRC is a domain-specific culinary corpus designed to capture language use in international recipe discourse. Specialized corpora typically exhibit a high concentration of topic-related terminology, resulting in lexical distributions that differ from those of general corpora. Culinary terms (e.g., *sauté*, *whisk*, *bake*, *dough*, and *pan*) occur with high frequency in recipe texts but appear infrequently in general-language corpora. Recipe discourse also emphasizes procedural clarity and practicality, leading to frequent use of imperative verbs (e.g., *stir*, *slice*, *chop*), ingredient names, and measurement expressions, whereas general corpora often prioritize narrative, descriptive, or argumentative language. These findings align with the multidimensional register framework proposed by Biber (1988), which argues that linguistic variation systematically reflects differences in communicative purpose, situational context, and discourse domain.

Additionally, culinary vocabulary reflects discipline-specific and cross-cultural content, which may not be represented in general corpora that emphasize broader American or global English usage. Differences in corpus construction procedures, annotation practices, and frequency-counting methods may also contribute to statistical variation. The observed lexical differences are expected outcomes of contrasts between specialized domain-focused corpora and broad general-language corpora, where domain specificity directly shapes vocabulary frequency and distribution. These findings are consistent with the register-variation perspective advanced by Biber & Finegan (1994), who argue that linguistic features systematically vary according to communicative function, genre conventions, and discourse community practices.

To examine their grammatical functional roles, the analysis shows that 790 items function as nouns, 188 as verbs, 140 as adjectives, and 34 as adverbs. A small number of lexical items display multifunctionality, including nine items functioning as both nouns and verbs and four as both verbs and adjectives. The findings indicate a strong predominance of nominal vocabulary in the corpus, highlighting the central role of ingredients, materials, and culinary entities in instructional recipe discourse. This distribution accords with the multidimensional register framework proposed by Biber (1988), which demonstrates that grammatical patterns systematically reflect communicative purpose and situational context, with informational and instructional registers typically exhibiting high nominal density.

When the compiled lexical items were compared with findings from previous related studies - namely Nordin et al. (2013), who generated a specialized word list for teaching food-related writing to Malaysian students, and Rungrueang et al. (2022), who analyzed food-service vocabulary - substantial overlap was observed. The comparison with Nordin et al. (2013) revealed forty-six shared lexical items, including terms related to ingredients, cooking processes, and food categories (e.g., *cook*, *garlic*, *pasta*, *meal*, *nutrition*, *vegetarian*, and *yeast etc.*). The comparison with Rungrueang et al. (2022) showed a larger overlap of 208 shared lexical items, encompassing

ingredients, cooking techniques, utensils, and culinary descriptors (e.g., *basil, blend, chop, garnish, marinade, oven, parsley, sauté,* and *whisk etc.*). Additionally, nineteen lexical items were found to overlap across all three studies, such as *chutney, garlic, ingredient, pasta, salad, stew, tofu, vegetarian, yeast, freeze,* and *fry*. These results indicate a strong convergence between the present lexical compilation and previously established culinary vocabulary lists, suggesting that a stable core set of high-utility food-related lexical items consistently appears across different learner groups, instructional contexts, and corpus sources.

The differences between the lexical items identified in the present study and those reported in Nordin, et al. (2013) and Rungrueang et al. (2022) can be attributed to several methodological and contextual factors. First, both Nordin et al. (2013) and Rungrueang et al. (2022) focused primarily on single-word lexical items, whereas the present study included both single-word items and multiword expressions, resulting in a larger vocabulary list. In addition, the earlier studies excluded words that overlapped with West's General Service Lists (West, 1953) and Coxhead's Academic Word List (Coxhead, 2000) in order to retain only highly specialized culinary terminology. In contrast, the present study intentionally retained overlapping items so that learners could access a comprehensive, all-in-one vocabulary resource representing frequently occurring lexical items across the entire cooking process. Second, the purpose and target users of the vocabulary lists differed. Nordin et al. (2013) developed a list designed for food-article-writing instruction, emphasizing academically and professionally relevant terms (e.g., *fermentation, hygiene, preservation*), while Rungrueang et al. (2022) focused on food-service communication, prioritizing operational and customer-service vocabulary (e.g., *order, serve, plate*). Consequently, some general cooking-process verbs (e.g., *mix, chop*) or specialized culinary techniques were less central to their objectives. Moreover, both lists were derived from specialized corpora reflecting specific cultural and institutional contexts (Malaysian instructional materials and food-service websites), which further influenced lexical selection. Third, corpus size played a crucial role. The corpus used by Nordin et al. (2013) contained approximately 3,698 tokens, while that of Rungrueang et al. (2022) included approximately 1.87 million tokens. In comparison, the present study analyzed a larger corpus of 7.09 million tokens drawn from one hundred cookbooks and over 2,000 recipes, more than 1,900 times larger than the corpus of Nordin et al. (2013) and nearly four times larger than that of Rungrueang et al. (2022). Consistent with corpus linguists such as Biber (1988) and Hyland (2008), larger corpora naturally yield a broader range of lexical items, which explains why the vocabulary list generated in the present study is more extensive than those reported in earlier research. The differences in lexical coverage are best explained by variations in lexical selection criteria, inclusion policies, research purposes, corpus composition, and corpus size, all of which influenced the resulting vocabulary inventories.

The classification of frequent lexical items in the SDUIRC demonstrates both similarities and differences when compared with the taxonomy proposed by Nordin, et al. (2013). While both studies share five comparable lexical domains - procedural terms, ingredients, utensils, descriptive terms, and other culinary-related vocabulary - the present study introduces an additional category: terms portraying the characteristics of culinary processes (e.g., adverbs describing how actions are performed). The absence of this category in Nordin et al. (2013) may be attributed to the small corpus size (3,698 tokens) and the nature of the instructional texts used in their dataset, where adverbial forms are considered stylistic and typically underused.

The six-category framework identified in the present study reflects the functional and communicative structure of recipe discourse, which requires explicit representation of procedures, ingredients, equipment, descriptive attributes, process modifiers, and supplementary contextual information. The prominence of terms associated with international culinary procedural aspects (193 items) confirms that recipes are fundamentally procedural texts characterized by imperative structures and sequential action verbs. The large number of terms linked to international culinary ingredients (464 items) reflects the global diversity of culinary traditions and the influence of globalization on ingredient diffusion. Similarly, the presence of terms related to international culinary apparatus (92 items) highlights the technological and cultural variability of cooking equipment across culinary traditions. The occurrence of terms delineating the attributes of international food and culinary apparatus (115 items) underscores the importance of precise sensory and qualitative descriptions - such as texture, color, and flavor - in guiding accurate food preparation. The additional category of terms portraying the characteristics of international culinary processes, and (34 items) emphasizes the communicative role of adverbial modifiers in enhancing procedural precision. Meanwhile, the other pertinent terms related to international food recipes (281 items) reflects the multidimensional nature of culinary discourse, encompassing nutrition, measurement units, meal contexts, cultural references, and food-service terminology.

An additional noteworthy finding concerns the high frequency of nationality-based food descriptors, with the most frequent being *Italian, French, Thai, Mexican, Indian, Greek, Japanese, Spanish, Chinese,* and *Korean*. The prominence of these nationality markers corresponds closely to global rankings of influential culinary traditions, indicating that the lexical distribution within the SDUIRC mirrors broader patterns of international culinary influence.

The lexical categorization results demonstrate that vocabulary distribution in international recipe discourse is strongly shaped by procedural communicative needs, globalization of ingredients and culinary techniques, technological developments in kitchen practices, and cross-cultural culinary exchange, all of which contribute to the distinctive linguistic profile observed in the corpus. These findings support the register-variation framework proposed by Biber (1988) and the discourse community perspective advanced by Hyland (2008), both of whom argue that linguistic patterns systematically reflect communicative purpose, disciplinary practice, and socio-cultural context.

The translation of the 1,165 culinary lexical items identified in the SDUIRC was carried out through the application of translation strategies proposed by Newmark (1998) and Baker (2011), with the aim of preserving meaning accuracy and cultural relevance. Strategies such as Transference, Naturalization, or Translation by using a loan word with explanation, sometimes combined with Classifiers or Superordinate (e.g., *springform* was translated as "pʰim sà.príŋ fɔ:m"), were used for terms without direct Thai equivalents. Cultural equivalent or Translation by cultural substitution were applied where culturally familiar terms improved comprehension (e.g., *coleslaw* was translated as "sà.làt kà.làm.plii", *currant* was translated as "lû:k kè:t"). When precise equivalents were unavailable, Paraphrasing or Translation by paraphrase using related word or unrelated word was employed to clarify meaning (e.g., *chorizo* was translated as "sâi.krɔ̀:k mŭ: sà.pen*"*; *guacamole* → krûaŋ jîm ʔa:.hă:n mék.sí.kàn). Translation by illustration was not employed in this study because the research did not aim to provide visual representations of the specialized words identified in the corpus. In addition, translation by omission was not used because the objective of the study was to present Thai equivalent meanings for all specialized words identified in the Suan Dusit University International Recipe Corpus. Omitting the translation of any lexical item would have resulted in an incomplete representation of the specialized vocabulary. Furthermore, the researchers considered every lexical item occurring in the recipe corpus to be important for conveying ingredients, procedures, and other technical aspects of bakery and pastry production; therefore, all lexical items were retained and translated.

Collocational analysis of food-processing verbs provides important insights into how procedural meaning is structured in culinary discourse. In corpus linguistics, collocation refers to the statistically significant tendency of lexical items to co-occur in predictable patterns, revealing conventionalized phraseological behavior that reflects disciplinary practices (Sinclair, 1991; Stubbs, 2001). Such recurrent combinations are not random but represent preferred lexical selections that encode domain-specific knowledge and typical action sequences. Within recipe discourse, verbs describing food preparation (e.g., *dice*, *chop*, *add*, *beat*) regularly occur with restricted sets of ingredients or materials, indicating specialized meaning associations that guide readers through step-by-step procedural instructions. Examining these collocations helps identify the operational vocabulary that characterizes culinary instructions and clarifies how linguistic regularities support clarity, efficiency, and standardization in instructional genres. For language learners, mastery of such verb–collocations enhances learning efficiency because frequently co-occurring combinations are processed and retrieved as integrated units rather than as isolated words. This facilitates more fluent comprehension and production of procedural instructions. Moreover, accurate use of established collocations reduces the risk of misinterpreting culinary actions, thereby contributing to procedural accuracy and successful task execution in professional and educational settings.

The findings suggest that international culinary discourse is characterized by a systematic lexical and collocational structure rather than a random collection of culinary terms. The identified lexical items can be organized into functional categories, while recurrent verb- collocations serve as the central mechanism linking ingredients, utensils, and procedural actions. Based on these findings, the study proposes the Lexical-Collocational Framework for International Culinary Discourse, consisting of three interconnected components: specialized lexical items, functional lexical categories, and procedural collocational patterns. This framework contributes to a broader understanding of how linguistic resources are systematically organized to construct culinary knowledge in international recipe discourse.

Suggestion

The researchers believe that this study can be further developed to analyze vocabulary in other types of recipes, various forms of writing, and types of examination question banks, such as those used in standardized English proficiency tests. Future research could investigate whether the use of linguistic corpora can aid in the development of English language skills, particularly in specialized English courses. Additionally, future studies might delve deeper into the differences in vocabulary usage between the recipes of one nation and those of another. Another promising avenue lies in the analysis of collocations in recipes, which would benefit students in their creative writing endeavors related to culinary arts. Furthermore, future research could compare the vocabulary commonly found in contemporary culinary corpora with that in historical culinary corpora to study diachronic changes in meaning and context. In this regard, researchers might consider using corpus analysis tools such as *AntConc* (Anthony, 2024), *Sketch Engine*

(Kilgarriff et al., 2014), or *#LancsBox* (Brezina & Platt, 2022) to manage and analyze large datasets. For historical corpora, innovative approaches to data compilation, such as digitizing vintage cookbooks or mining archival materials, could be explored. To align with broader interdisciplinary perspectives, future research could also engage with fields such as cultural studies to examine the socio-cultural implications of culinary vocabulary, or artificial intelligence to utilize machine learning for deeper semantic analysis. Investigating the role of visual elements (e.g., food photography, illustrated instructions) in shaping culinary language, or examining the impact of digital platforms (e.g., social media, cooking blogs, and influencer-generated content) on emerging culinary vocabulary, could offer fresh and forward-thinking directions. These perspectives would not only enrich linguistic and educational insights but also expand the academic reach to encompass media studies, digital literacy, and culinary education. These proposed directions align with the broader socio-cultural discourse approaches associated with Fairclough (1992) which supports the expansion of corpus-based culinary research across genres, cultures, historical periods, and digital environments.

Beyond its pedagogical implications, this study contributes theoretically to corpus linguistics by demonstrating how lexical frequency, statistical validation, functional classification, and collocational patterning can be integrated within a unified corpus-driven framework to describe domain-specific discourse. The findings extend research in English for Specific Purposes (ESP) by providing empirical evidence that specialized culinary discourse is structured not only by frequency-based lexical selection but also by systematic phraseological patterning. In doing so, the study advances methodological approaches to identifying and validating domain-specific lexis in professional genres. These pedagogical implications are consistent with the ESP genre-based approach advocated by Hyland (2008), the frequency-based vocabulary learning principles proposed by Nation (2001), and corpus-informed pedagogical perspectives discussed by Hunston (2002), all of which emphasize the importance of authentic frequency patterns and phraseological structure in specialized language instruction. This research contribution is in line with the corpus-driven principles by Sinclair (1991), which emphasizes the systematic relationship between lexical frequency, phraseological patterning, and professional discourse structure.

Several limitations should be acknowledged. First, although the SDUIRC is substantial in size, it represents a single genre restricted to English-language recipe books and excludes other culinary discourse types such as restaurant reviews, instructional videos, or professional manuals. Second, the reliance on highly downloaded digital texts may introduce platform-related or regional biases. Third, while cross-corpus comparisons support domain specificity, generalizability beyond recipe discourse should be interpreted with caution. Future research may therefore extend the corpus to additional genres and multimodal formats to enhance representativeness.

In summary, this study is novel in its integration of large-scale corpus compilation, multi-criteria statistical validation, functional lexical classification, and systematic collocational analysis within a single investigation of international culinary discourse. By identifying and validating 1,165 domain-specific lexical items and their phraseological patterns, the research provides a replicable methodological model for future corpus-based studies of professional and instructional genres. The findings not only support corpus-informed ESP development but also establish a foundation for expanding culinary discourse research across languages, cultures, and historical periods.

Author Contributions

Woravit Kitjaroenpaiboon: Conceptualization; Methodology; Investigation; Data curation; Formal analysis; Validation; Visualization; Writing – original draft; Writing – review & editing; Project administration; Supervision; Funding acquisition
Sirote Pholpunthin: Conceptualization
Samniang Fakrajang: Validation
Niphatchanok Najpinij: Correspondence with the journal
Prissana Fongsrun: Validation
Jareuk Sriaroon: Validation
All authors have read and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

ORCID iD

Woravit Kitjaroenpaiboon	https://orrcid.org/0009-0003-5166-3736
Sirote Pholpuntin	https://orrcid.org/0009-0006-4522-785X
Samniang Fahkrajang	https://orrcid.org/0009-006-3172-00045
Niphatchanok Najpinij	https://orcid.org/009-002-5265-6696
Prissana Fongsrun	https://orrcid.org/0009-0001-7026-7769
Jareuk Sriaroon	https://orrcid.org/0009-0006-9597-5637

Conflict of Interest

The authors declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of the use of Generative AI or AI-assisted tools

During the preparation of this manuscript, the authors used ChatGPT (OpenAI) to assist with improving the clarity, grammar, academic style, and organization of the manuscript. All AI-assisted outputs were critically reviewed, edited, and verified by the authors. The study design, corpus compilation, data collection, statistical analyses, interpretation of the findings, and conclusions were conducted solely by the authors, who assume full responsibility for the accuracy, originality, and integrity of the final manuscript.

References

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

Aroonmanakun, W. (2011). Corpora and emerging technology for ELT. *Journal of Studies in the English Language, 6*, 221–236.

Baker, M. (2011). *In other words: A coursebook on translation* (2nd ed.). Routledge.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

Biber, D., & Finegan, E. (1994). *Sociolinguistic perspectives on register*. Oxford University Press.

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics, 20*(2), 139–173. https://doi.org/10.1075/ijcl.20.2.01bre

Brezina, V., & Platt, W. (2022). *#LancsBox X* [Computer software]. ESRC Centre for Corpus Approaches to Social Science (CASS). https://lancsbox.lancs.ac.uk/

Chung, T. M., & Nation, P. (2004). Identifying technical vocabulary. *System, 32*(2), 251–263. https://doi.org/10.1016/j.system.2003.11.008

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics, 32*, 45–61. https://doi.org/10.1017/S0267190512000074

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

Dang, T. N. Y., Webb, S., & Coxhead, A. (2022). Evaluating lists of high-frequency words: Teachers' and learners' perspectives. *Language Teaching Research, 26*(4), 617–641. https://doi.org/10.1177/1362168820911189

Fairclough, N. (1992). *Discourse and social change*. Polity Press.

Galindo-Cuspinera, V., & Breslin, P. A. S. (2006). The liaison of sweet and savory. *Chemical Senses, 31*(3), 221–225. https://doi.org/10.1093/chemse/bjj022

Haspelmath, M. (2012). How to compare major word-classes across the world's languages. *UCLA Working Papers in Linguistics, 17*, 109–130. https://doi.org/10.5281/zenodo.3678496

Hunston, S. (2002). *Corpora in applied linguistics.* Cambridge University Press. https://doi.org/10.1017/CBO9781139524773

Hyland, K. (2008). Genre and academic writing in the disciplines. *Language Teaching, 41*(4), 543–562. https://doi.org/10.1017/S0261444808005235

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography, 1*, 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kitjaroenpaiboon, W., & Getkham, K. (2015). An analysis of interactional metadiscourse devices in communication arts research articles. *International Journal of Management and Applied Science (IJMAS), 1*(9), 125–131.

Kitjaroenpaiboon, W., & Getkham, K. (2016). Stylistic patterns in language teaching research articles: A multidimensional analysis. *PASAA: Journal of Language Teaching and Learning in Thailand, 52*, 169–208.

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

Kitjaroenpaiboon, W., Pholpuntin, S., Fahkrajang, S., Najpinij, N., Fongsrun, P., & Sriaroon, J. (2025). *An analysis of international food processing lexis and formulaic languages: An application of a corpus linguistic based approach.* [Unpublished research report]. Suan Dusit University.

Kitjaroenpaiboon, W., Wongwiseskul, S., Puksa, T., & Khamsakul, B. (2021). 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.

Laosrirattanachai, P., & Ruangjaroon, S. (2020). The word lists of hospitality service review construction. *Journal of Studies in the English Language, 15*(1), 107–158. https://so04.tci-thaijo.org/index.php/jsel/article/view/211119

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*(140), 5–55.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

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. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Newmark, P. (1998). *More paragraphs on translation*. Multilingual Matters.

Nini, A. (2019). The Multi-Dimensional Analysis Tagger. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis: Research methods and current issues* (pp. 67–94). Bloomsbury Academic.

Nordin, N. R. M., 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

Rashid, M. H., Lan, Y., & Hui, W. (2022). The importance of vocabulary in teaching and learning in applied linguistics. *Linguistics and Culture Review, 6*(S2), 541–550. https://doi.org/10.21744/lingcure.v6nS2.2177

Ricketts, J., Nation, K., & Bishop, D. V. M. (2007). Vocabulary is important for some, but not all reading skills. *Scientific Studies of Reading, 11*(3), 235–257. https://doi.org/10.1080/10888430701344306

Risteski, M., & Korunovski, S. (2016). Culture and food [Paper presentation]. XVI International Conference of Partner Institutions "Planning for Change," Ohrid, Republic of Macedonia.

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.

Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics.* Blackwell.

Sudarmo, S. (2021). The importance of speaking in English as a foreign language between skillful and thoughtful competencies: Studying sociolinguistics perspectives. *Linguistics and Culture Review, 5*(S1), 113–124. https://doi.org/10.21744/lingcure.v5nS1.1321

Suwonsichon, S. (2019). The importance of sensory lexicons for research and development of food products. *Foods, 8*(1), 27. https://doi.org/10.3390/foods8010027

Tenopir, C., Volentine, R., & King, D. W. (2012). Scholarly reading and the value of academic library collections: Results of a study in six UK universities. *Insights, 25*(2), 130–149. https://doi.org/10.1629/2048-7754.25.2.130

Wang, Q. J., Mielby, L. A., Junge, J. Y., Bertelsen, A. S., Kidmose, U., Spence, C., & Byrne, D. V. (2019). The role of intrinsic and extrinsic sensory factors in sweetness perception of food and beverages: A review. *Foods, 8*(6), 211. https://doi.org/10.3390/foods8060211

West, M. (1953). *A general service list of English words: With semantic frequencies and a supplementary word-list for the writing of popular science and technology.* Longmans, Green.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press. https://doi.org/10.1017/CBO9780511519772

Article info
Received: 5 August 2024
Revised: 1 July 2026
Accepted: 11 August 2026

Authors
Kitjaroenpaiboon et al.

Email
woravit_kit@dusit.ac.th