TOWARD AN UNDERSTANDING OF PHRASEOLOGICAL UNITS AND PRAGMATIC FUNCTIONS IN RESEARCH ARTICLES

ATTAPOL KHAMKHIEN

Centre for Applied Linguistic, University of Warwick
E-mail: a.khamkhien@warwick.ac.uk

Abstract- This paper presents a combination of corpus-based approach and qualitative analysis in creating a list of phrase logical units useful for writing for publication. A corpus of research articles published in journals indexed in the Thai Citation Index (TCI) database, consisting of the four internal sections of these articles (approximately 205,187 running words) was compiled. The Sketch Engine was used to identify n-grams in the corpus, and three-grams appeared to be pedagogic useful according to their grammatical and semantic relations. Five EAP instructors were invited to select these three word n-grams, contributing to a pedagogically useful list of 289 phrase logical units. The findings indicate that most of the phrase logical patterns primarily performed as ‘research-oriented bundles’, followed by ‘text-oriented, stance-oriented bundles, and others category’. Pedagogically, this study sheds light on language use in this research genre and provides more evidence-based instructional practices, especially for advanced language courses targeted at scholarly publications.

Index terms- Phraseology, Academic word list, Corpus, Research articles, TCI database.

I. INTRODUCTION

It has become fairly well established that phraseological units can facilitate second language speakers to learn a language. They may help speakers appear more nativelike, as they provide ready-made chunks of language [1] which are appropriate to specific contexts. Pedagogic implications can be found, in writing academic works and in several areas related to academic communication [2] [3] [4] [5] [6] [7] [8]. Many researchers have assumed that the use of fixed expressions represents fluent linguistic production in both spoken language [9], and written or academic texts [3] [6] [10]. Phraseological patterns are defined in certain ways. In essence, they are combinations of three or more words that are identified in a corpus of natural language by means of corpus analysis software programs. Researchers may use different terms referring to studies of phraseology. For instance, Altenberg [11] preferred to use the term “recurrent word-combinations” in investigating word patterns typically recurring in spoken English. The term “lexical bundles” was first defined in 1999 by Biber and his colleagues to describe the patterns of “words that show a statistical tendency to co-occur” [10]. Erman and Warren [12] stated that they were “combinations of at least two words favored by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization”. Based on our intuition, the word units considered to be lexical bundles must be unrecognized by language speakers and researchers. Fixed expressions or so-called idiomatic phrases that have fixed meanings and can be understood by language speakers cannot be included because lexical bundles are distinct from those [13].

To identify multiword combinations, a frequency-based method is most appropriate for large corpora of hundreds of thousands of words. In fact, researchers set certain criteria before scanning and analyzing a corpus, which cutoffs can generally range from 10 to 40 occurrences per million words [10] [14]. However, it has some limitations for use with small data sets, and using frequency alone as a criterion for formulaicity might not give any information about the psycholinguistic validity of the formula [1]. It may yield meaningless combinations of words for functional analyses.

In this regard, for Thai novice researchers and graduate students, writing a research article and getting it accepted for publishing in a peer-reviewed journal are vital. The task, however, may be difficult for them as article writers are expected to use language which is appropriate to their discipline and writing convention. Since phraseology has been claimed that it facilitates the development of long-term language ability, and permits a learner to more easily use frequent word units for meaning reference, two specific objectives of the present study are: 1) to identify phraseological patterns appear in research articles indexed in the TCI database, and 2) to classify the phraseological patterns generated with pragmatic functions. It is expected that this study can provide a better understanding of the use of phraseological patterns in writing for scholarly publications, especially in the journals in the TCI database. The findings can be incorporated into an advanced English course focusing on reading and writing research articles.

II. RELATED STUDIES

A number of studies of formulaic language have been carried out to identify and retrieve multiform units for linguistic research and language teaching and learning [4] [5] [6] [10] [14]. Jalali, Rashek and Rizi [15], for example, explored possible differences and similarities in reported usage of bundles by professional writers and graduate students in...
academic writing. Data of 1,217,963 words were taken from two collections of academic texts. Structural categories proposed by Biber et al. [10] and Hyland’s [7] comprehensive taxonomy were adopted to identify functional categorization of the bundles. The sequences were shown that the frequency, structure and function of lexical bundles between students’ theses and dissertations and articles were different, even though the corpus was from the same discipline. Published academics were found to be more dependent on phrasal rather than clausal bundles. At the functional level, research-oriented bundles were prevalent in both articles and theses. Cortes [6] investigated the relationship between lexical bundles and rhetorical patterns captured by move-based genre analysis in 1,372 Introduction sections of research articles taken from 56 journals available in the Published Research Article Corpus (PRAC) from various disciplines, including soft and hard sciences. Using a variation of the Lexical Bundle Program (LBP), Cortes reported that a group of lexical bundles were exclusively linked to one move or a step in a move, while some lexical bundles could occur across moves and steps. Working to construct the Academic Word List (AWL), Coxhead [16] selected word families on the ground they occurred 100 times or more in a corpus of 3,500,000 words taken from journal articles and textbooks in four different areas. This list is aimed for advanced EFL learners at the tertiary level. Nonetheless, since the list is based on the notions of frequency and range, in a recent study, Simpson-Vlach and Ellis [14] created an Academic Formulas list (AFL) by adding elements of native speaker judgment to the process of identifying lexical bundles as they note that many of the items identifies by lexical bundle research are of limited utility for the purposes of teaching learners to be competent with academic discourse. EAP instructors were invited to rate the formulas to determine whether the phases found were a formulaic expression, a phrase, or expression. All of the processes made it possible to select multiword formulas and three functional categories were reached: referential expressions, stance expressions, and discourse organizers as suggested by Biber et al. [17]. Contrast and comparison category was added into referential expressions group, and the discourse organizer category was modified. Similarly, Martínez and Schmitt [18] combined frequency and qualitative criteria in selecting individual words and phrasal expressions from the BNC corpus. Exhaustively, a random sampling technique was applied to the process of searching derived multiword lexical items line-by-line to check whether these items were phraseological polysemy. The final list called the PHRASE List ended up, consisting of 505 phraseological items. The list was claimed to be “useful for pedagogic materials including more multiword items, such as textbooks, graded readers, and language tests” (p. 316). However, the usefulness and benefits of applying the list are still open to question because, similar to Simpson-Vlach and Ellis [14], the functions of these multiword items were not provided, which might at first seem difficult, especially for lower-proficiency learners.

The above studies are well represented by description derived from textual investigations and comparisons of corpora, including specialized and available corpora, and native and non-native speakers corpora, etc. These studies, however, have yielded a distinctive list of phraseological patterns due to the different corpora or texts analysed. Moreover, functional analysis of the word combinations is essential to their value as teaching items. It is considered to be both important and essential to explore the use of phraseological patterns in the articles, as it serves as a useful guide in drafting an article. The result of this line of research would provide further evidence to facilitate a better understanding of language usage and the use of phraseological patterns in writing articles.

III. METHODOLOGY

Compiling the corpus
A corpus of 50 research articles taken from nine journals was complied. All the selected journals were classified in tier 1 in the TCI database, which would be further included in the ASEAN Citation Index [19], representing the high quality of the journal. To control for potential changes within the discipline and across time and to enhance the validity of the results of the study, the sampling of the journals was restricted to the years 2013 and 2014 only. It is acknowledged that some other factors, e.g., style of writing used in the articles, peer-review process and copy editing process, would remain in these selected articles. Given the focus of the present study on the four internal sections of the articles (Introduction, Methods, Results, Discussion), other sections of the articles were excluded from the analyses. These systematic procedures yielded a corpus of approximately 205,187 running words. All of the systematically selected 50 articles were converted into text files and incorporated into the SketchEngine [20] to be used for the analysis.

Processing and identifying n-grams
The wordlist option in the SketchEngine was used to explore two-, three-and four-word n-grams which refer to as high frequency formulaic expressions in the corpus. To identify phraseological units based solely on frequency of occurrence, cut-offs of frequency are somewhat arbitrary [7]. The frequency threshold was set with each frequent n-gram reported occurring a minimum of five times in the corpus. In addition, to guard against the subjectivity and idiosyncratic expressions introduced by individual
writers, all the n-grams generated were carefully checked to ensure that they occurred in at least three files in the corpus, representing the occurrence of such n-grams in at least three articles. Next, based on the length and number of words of the corpus, the choice of norming to 1,000 words was appropriate to use in the present study [10] to make comparisons of the findings of the study with previous studies possible.

Even though four-word bundles are claimed to be more phrasal in nature [2] [4] [6] [17] [21] [22], in the present study, a list of two-word, and four-word n-grams appeared ungrammatically complete (e.g. of the, in the, to the), and content-based lexical items in relation to a particular subject-matter (e.g. intrinsic motivation of English, in the English language), which was simply an artefact of what the writers were writing about and “neither terribly functional nor pedagogically compelling” [14]. The three-word n-grams, in contrast, appeared to be more interesting, because many of them constituted complete syntactic units as being independent meaningful phrases, including a number of grammatical items expressing semantic relations (e.g. in order to, as well as). Even if most of them do not represent complete structural units (e.g. the use of, the results of), they are still seen as “important building blocks in discourse” [2].

A set of qualitative criteria was applied by first removing content-based strings or noun groups as they were considered not pedagogically interesting. A first list of 476 potential n-grams was obtained. I went down the n-gram list item-by-item, looking for ‘plausibly formulaic’ multiword strings [13] which realized pragmatic functions. To assure a high degree of reliability of the list, five EAP-experienced English instructors were invited to select the items that appeared to be pedagogically useful for research reading and writing. They were advised that the purpose of the exercise was the construction of a list of phraseological patterns worth learning and teaching when drafting an article for publication. Each of the potential three word n-grams carefully chosen by at least three instructors was included in the final list. The selected phraseological patterns were further explored in order to investigate to what extent they were used by article writers. They were then functionally categorized according to their pragmatic meanings.

IV. RESULTS

All of 289 phraseological patterns found pedagogically interesting were grouped into functional categories, based on those of Biber et al. [10] and Hyland [7]. A combination of these two categories was chosen because it was more appropriate for the phraseological patterns identified. The following Restable presents examples of phraseological united identified with the frequency of each type of pragmatic functions.

<table>
<thead>
<tr>
<th>Pragmatic functions</th>
<th>Examples of phraseological patterns identified</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research-oriented</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>at the beginning, from this study, the current study, the present study, this study is</td>
<td>12</td>
</tr>
<tr>
<td>Procedure</td>
<td>an analysis of, an investigation of, can be used, data were analysed, in order to, investigation of the, the participants were, the use of, was carried out, was divided into, was used to, were asked to, were used in</td>
<td>52</td>
</tr>
<tr>
<td>Quantification</td>
<td>a corpus of, a lot of, a number of, a part of, A total of, a variety of, all of the, each of the, the large number of, majority of, the most of, the number of, one of the, some of the, the frequency of, the level of, the majority of, the number of criteria based on, the meaning of, the pattern, the study of</td>
<td>26</td>
</tr>
<tr>
<td>Description</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Intangible framing</td>
<td>an important role, reliability of the, the development of, the importance of, the influence of, the role of</td>
<td>29</td>
</tr>
</tbody>
</table>

| Text-oriented       |                                               |       |
| Transition signals  | as a result, as well as, in addition to, in line with, in other words | 10    |
| Resutative signals  | a result of, agreed that the, findings show that, have shown that, indicate that the, is consistent with, point out that, result shows that, results of the, revealed that the, the findings from, the findings of, the result of, The results show, This indicates that | 58    |
| Structuring signals | above table, showed, according to the, they are presented in, are shown in, As shown in, based on the, below illustrates the, presented in Table, seen in Table, shown in Table, The above table | 22    |
| Framing signals     | As for the, for further research, from the context, in relation to, in terms of, in the future, terms of the, the basis of, the context of, the other hand, the part of, the purpose of, with regard to | 17    |
| Stance-oriented     | are likely to, are more likely, because of the, can be seen, compared to the, considered as, a contribute to the, due to the, is possible that, is suggested that, it can be, it could be, it is possible, might not be, more likely to, seems to be, should be noted, This is because | 45    |
| Engagement Others   | developed by the, exist at the, fact that the, identified according to, study aims to, the fact that, the sense that | 11    |

Research-oriented bundles

Research-oriented bundles serve to help writers structure their activities and experiences, which include those referring to research location or place,
procedures, quantification, description and topic of the research, and intangible framing attributes which indicate an abstract property of something.

(1) The instrument employed in this study was an LLS questionnaire. It focused on two sections: background information and SILL. [LEA 1]

(2) In order to test the reliability of the scoring system, a second rater was asked to recheck the possibility of the results. [JES 1]

(3) This is because most of the students would normally choose the statement that contained similar words to the main statement, which were ‘park’ and ‘parked’. [SPPJ 1]

(4) The results of this study confirmed to the study of (Ref). [LEA 6]

(5) Great attention is paid to grammatical aspects, whereas the importance of collocations is overlooked; likewise, lexical choices in a second language structure are often arbitrarily and improperly produced. [LEA 8]

In (1), in this study, allows writers to prepare readers for the information about the research samples and instruments used in the study. Reader preparation is achieved by the use of a demonstrative (this), allowing the writers to frame the ‘location’ and pinpoint specific information of the research methodology of the study. Meanwhile, in order to, in (2) indicates one of the ‘research procedures’ accomplished in the study. It can be used to state how or why something is done/what something is for [23].

Another common type of research-oriented bundles is ‘quantifying expressions’ (most of the), as shown in (3), qualifying a proposition with an expression relating to anything potentially measurable [21]. In (4), the study of, indicates the physical properties of the study being discussed. Likewise, the importance of in (5) is an example of ‘intangible framing attributes’, indicating an abstract property of something. It is used to highlight the importance of the topic being presented, alerting readers to understand the need for conducting the present study.

**Text-oriented bundles**

Text-oriented bundles are united by a broad function of showing relationships, and a review of the distinctive bundles in this category shows the two bundles focusing on relationships of rather different sorts [7] [23]. This group includes the transition, resultative signals and framing signals of the text.

(6) This was explained later in the interview, in which the respondent acknowledged that the strategy instruction helped him what strategies to use and how to use them; as a result, it improved his reading comprehension. [JLA 1]

(7) The scores gained for each group showed that the Thai version favored the low-score group the most, followed by the mid-score group, with the high-score group not benefiting at all. [JES 1]

(8) The analysis of the findings shows that Thai EFL students have positive attitudes towards learning to speak English in the CLT classroom in Australia. [NET 4]

(9) Figure 1 below illustrates the complete vocabulary learning strategy classification found from the stage of the data analysis. [NET 12]

(10) At this point, the survey presented in this paper is preliminary, and, thus, the perspectives for further research are vast, and, as such, will be improved and expanded on in further works. [JES 2]

The grammatical phrase as a result is often used and functions as a conjunctive adverbial phrase to indicate cause-and-effect relationship. As in (6), this string is considered as a ‘transition signal’, which links a logical idea with the preceding sentence. The strings showed that, in (7), and finding shows that, in (8), as ‘resultative signal’, denote an integration of reporting verbs and past tense when research findings are presented. Definite articles (the) in the text indicate that the shared knowledge between the writers and the readers is established based on the preceding context [10]. Again, below illustrates the, in (9), is used to report findings and, at the same time, helps refer readers to other parts of the text, signifying graphical information. Based upon Hyland’s [7] functional classification, this type of cluster can be called ‘structuring signals’. In (10), the string for further research is used as a ‘framing signal’, indicating that, within the context, the statements of findings remain to be substantiated or validated by future studies.

**Stance-oriented bundles**

The writers explicitly convey epistemic and affective judgements, evaluations and degrees of commitment to their claims with stance-oriented bundles. These bundles include writers-and reader-focused features of discourse. (11) It is possible that the Thai learners in our study were led by the distributional bias in the input, thus producing simple past tense forms with state verbs. [MNY 5]

(12) All of them are intertwined to promote student engagement, which seems to be a source of effectiveness in teaching a foreign language. [JLA 2]

(13) However, it should be noted that there were significant differences in the level of attention/awareness between the different groups, whereas there were no differences in the frequency of actions they took in response to the type of written corrective feedback they had received. [NET 3]

As can be seen from the preceding instance, it is in agreement with Hyland’s [7] explanations in that the use of anticipatory it structures deals with the writer’s preference for the impersonality in the text. The strings is possible that and seems to be in (11) and (12) are often used to express a writer’s evaluation of a proposition with regard to uncertainty, epistemic standing or modality. These phrases also convey the writer’s attitude about proposition, allowing him or her to present information as an opinion rather than as accredited fact [7]. In (13), the phrase it should be
noted that (engagement feature) expresses the writer’s claim without explicitly identifying its source.

Concerning the other categories, it is commonly acknowledged that there is no context-free correspondence between structural patterns and pragmatic functions, because the meaning of an utterance in a language can vary widely according to its use in particular context, interpretation, sociocultural factors, social conventions, etc. [24]. This group, thus, includes those that express some pragmatic functions depending on the contextual environment (e.g. developed by the, fact that the, study aims to).

**DISCUSSION**

As shown by the findings of the current study, a combination of Hyland’s [7] functional classification and Biber et al.’s [17] taxonomies was appropriate and useful, revealing characteristics and pragmatic functions of the phraseological patterns identified, including research-oriented, text-oriented and stance-oriented functions. Even though this study focuses on three-word n-grams, which is distinctive from previous studies investigating four-word bundles, most of the findings presented are partly in accordance with several previous studies [7] [17] [23] [25]. A possible reason explaining this finding is that Biber et al.’s taxonomy was derived from analyzing a huge corpus, including a mixture of various disciplines and registers, including written and spoken texts, while Hyland’s classification was designed specifically for lexical units found in academic writing. A combination of these and an inductive approach, thus, nicely fits the phraseological patterns found in the corpus of articles in this study. The present study also shows that a combination of a corpus-based approach and qualitative analysis is helpful in identifying phraseological units since they are not necessarily formulaic just because they occur together often in a corpus of specific contexts. The results generated by this study, thus, have generated a number of important implications for graduate students and novice writers in reading and drafting articles. The pragmatic functions of phraseological patterns identified can serve as a reference for raising writers’ awareness when interpreting underlying meanings of academic texts and materials in general, particularly articles published in the TCI database. Showing how a pragmatic function can be operationalized to allow for creating meanings in different situational contexts is in response to Biber et al. [26], who succinctly noted that corpus-based analyses must go beyond simple counts of linguistic features, but that it is crucial to include qualitative, functional interpretations of quantitative patterns. I believe that these approaches are not desirable, as the process is time-consuming, but can, to some extent, be fruitful, particularly with the genre-specific corpus used in this study, in facilitating the task of article writing. Pedagogically, the knowledge obtained from the current study may serve as a useful basis for preparing manuscript drafts for journal submission. Instructors can integrate knowledge from this study with the use of authentic texts to demonstrate how phraseological patterns are frequently used, as well as the functions they perform in context. In other words, graduate students should be encouraged to associate those functions and forms within the context of language use. The present study has a contribution to awareness-raising of the functions of repeated patterned-expressions in academic texts, and how they may enhance students’ overall expertise in writing English for publication. Material designers and practitioners can consider a combination of quantitative and qualitative aspects in selecting useful phrases, including the identification of their pragmatic functions, as a reference in developing textbooks and classroom materials.

**CONCLUSION**

The present study set out to investigate the use of phraseological patterns and their functions in applied linguistics research articles. As generated by the SketchEngine, unlike previous studies investigating four-word bundles [3] [4] [6] [7] [10] [22] [25], the current study specifically explored the quantitative distribution of three-word n-grams in the corpus of the selected research articles, as they were considered psycholinguistically unproblematic and pedagogically interesting [14]. This paper has also outlined the methodological processes and important considerations, including qualitative perspectives and judgments from EAP instructors, rather than simply extracting frequent words from a corpus in selecting a list of three-word n-grams. The methods used here could, thus, ensure that the phraseological patterns generated are pedagogically motivated and merit the pedagogic purposes. Bringing all of those strands together — quantitative frequency, and qualitative judgments about what are meaningful phrases, and inputs from experts in the field to consider those useful phrases - demonstrates a thorough perspective on conducting textual analysis to receive in-depth and specific information, verify the outcomes and ensure that the results of this study are accurate and reliable.

**REFERENCES**

Toward An Understanding of Phraseological Units and Pragmatic Functions In Research Articles


★★★★