Manuscript version: Published Version
The version presented in WRAP is the published version (Version of Record).

Persistent WRAP URL:
http://wrap.warwick.ac.uk/145707

How to cite:
The repository item page linked to above, will contain details on accessing citation guidance from the publisher.

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher’s statement:
Please refer to the repository item page, publisher’s statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk
Constructing subject-specific lists of multiword combinations for EAP: A case study

Abstract: This study combines a corpus-based approach and intuition-based judgements to develop a set of multiword combinations for research publications in academic journals. To obtain a representative sample, a corpus of four internal sections of 120 Applied Linguistics research articles indexed in the TCI (Thai Citation Index) database was systematically compiled and investigated. To identify n-grams which occur frequently in the corpus, a corpus-based approach was used. First, a list of 49 content-based strings, likely to be the most useful for pedagogic purposes, was derived. Based on their grammatical and semantic relationships, 3-grams were further investigated. For multiword sequences to occur frequently in the corpus, some pragmatic functionality is required which contributes to pedagogical use. Five EAP instructors were therefore invited to select the useful multiword combinations from the list of identified n-grams. A list of 289 phraseological patterns was finally created successfully. The list can provide additional evidence-based and corpus-informed instructional resources which support English teachers with the planning of lessons as well as materials design and development, particularly for advanced language courses which target scholarly writing.

Keywords: academic word list; research articles; phraseology; English for Academic Purposes (EAP); writing for publication

1 Introduction

Composing a research article can be a formidable task for a graduate student or novice writer, since writing competency for publications needs a background knowledge of the genre and its associated textual features (Feak and Swales 2011;
Hyland 2004). An article writer must have awareness of the generic and microstructures, macro-structures, and linguistic features of articles in the academic genre (Hyland 2008a, Hyland 2008b; Samraj 2005; Swales 1990, Swales 2004). In terms of micro-structures, vocabulary is one of the challenges for L2 and EFL learners when taking part in reading and writing academic discourse (Shaw 1991). In Thailand, novice scholars and graduate students feel an increased pressure to publish their research work in prestigious journals, especially those indexed in the TCI (Thai Citation Index) database. Most universities require graduate students to publish a minimum of one research article in a peer-reviewed journal as a component of graduation. To assist people in successfully publishing their research in a peer-reviewed journal, extra attention must be paid to the vocabulary in an English for Academic Purposes (EAP) context, particularly the specialist words which frequently occur in research articles.

Having knowledge of vocabulary is seen as an important part of learning a language. It influences the reading and writing proficiency of the learner (Nation 2001) and is strongly linked with linguistic ability and academic achievement (Jacobs 2008; Gardner and Davies 2014). Since acquiring vocabulary has a crucial role in language learning, researchers have paid attention to the academic vocabulary used in the literature in order to help language learners to achieve their academic goals. Several academic word lists have been examined in different academic genres and specific disciplines, e.g. Tangpijaikul’s (2014) frequent word lists in economics and business; Watson Todd’s (2017) opaque words list in engineering; Coxhead’s (2000), Gardner and Davies’s (2014), and Simpson-Vlach and Ellis’s (2010) frequently-used words lists in general English; Yang’s (2015) Nursing Academic Word List (NAWL); Brezina and Gablasova’s (2015) general vocabulary list representing current language use; and Lei and Liu’s (2016) list of frequent words required for the medical learner. Such academic word lists have been developed from various subject-specific aspects; however, the main objective of vocabulary lists is generally to meet the needs of language learners and build decision-making tools for EAP and ESP instructors in regard to teaching, learning, and material and curriculum design.

Some of the academic word lists, e.g. the AWL by Coxhead (2000) and the General Service List (GSL) by West (1953) have received criticism since they included lists of word families often used in the English language. Nation (2001) stated that more significant words exist than appear in the lists, but since they are not used as often in data analysis, they do not show up in the lists. Academic vocabulary lists in the literature have been investigated via various methods and were created from various corpora across academic disciplines (Gardner and Davies 2014). Hyland and Tse (2005) argued how useful a general academic vocabulary list can be, e.g. the AWL by Coxhead (2000), since vocabulary can
vary across disciplines in terms of frequency, range, collocation, and meaning. Furthermore, a number of the vocabulary lists provide individual vocabulary items which were distributed in a different frequency order (such as AWL and GSL) with regard to the principles on which the lists were created as well as with respect to their utility and the researchers’ purposes. In addition, because the majority of academic word lists in past research includes only single words and word families, e.g. benefit, beneficial, beneficiary, beneficiaries, benefited, befitting, and benefits, which could be used across a variety of occurrences and contexts (Brezina and Gablasova 2015), such lists have some limitations.

Aside from single word vocabulary lists, multiword combinations are said to help as they offer a “pre-packaging of information or of the structures used to present information” (Reppen 2004: 83), which assists the writer by lowering the processing load. As a result, learning and teaching language can be achieved in the writing of academic works and in related areas of academic communication (e.g. Biber and Barbieri 2007; Conrad and Biber 2004; Cortes 2004, Cortes 2006, Cortes 2013; Hyland 2008a, Hyland 2008b; Li and Schmitt 2009; Martinez and Schmitt 2012). There is a general consensus that such knowledge plays a facilitating role in the learning and use of a language as it represents fluent linguistic production, especially in spoken language (Pawley and Syder 1983) and academic texts (e.g. Biber et al. 1999; Conrad and Biber 2004; Cortes 2013; Hyland 2008a, Hyland 2008b; Nesi and Basturkmen 2006). Thus, it is interesting and important to develop new lists of academic words including multiword combinations which are derived from new methods, aside from using frequently used as a selection criterion, for example range and frequency, and from a particular or specific discipline.

Because of the need of Thai graduate and novice students to publish research articles in English, the principal objective of this research is to show the use of newer methods to develop multiword combination lists covering essential and useful words for publication in applied linguistic research. We expect that the creation of discipline-specific word lists will be important and improve the academic skills of graduate students regarding the writing and reading of English research articles.

2 Related literature

Academic word lists

To assist language learners in developing their knowledge of vocabulary, West (1953), a leading pioneer in the field, developed the General Service List of 2000.
word families chosen from a 5-million word corpus. Although West in developing
the GSL used various criteria, including frequency, ease of learning and covering
useful ideas, as well as stylistic level and emotional neutrality (West 1953: ix–x),
the list has been criticized in regard to age and the number of words included.
Gilner and Morales (2008) also question the possibility of expanding the GSL
given the combination of objective and subjective criteria on which the original
word list was based. Some words in the GSL, however, may occur less frequently
in specific other fields (Xue and Nation 1984). Therefore, Xue and Nation (1984)
decided to create the University Word List (UWL) by adding high-frequency, non-
overlapping words to Lynn’s (1973) and Ghadessy’s (1979) word lists. The UWL
consists of 737 base words and is stated to be a useful and complete resource for
language students.

Coxhead (2000) argued that the UWL did not have consistent selection prin-
ciples and was derived from small corpora, failing to cover a broad variety of
topics. She created an Academic Word List (AWL) which contains 570 word fami-
lies from a corpus of 3.5 million over four different areas – natural science, law,
commerce, and arts. Each word in the AWL list must occur over ten times in each
of the disciplines, in at least 15 of the 28 subject areas, and over 100 times in the
entire corpus. The AWL is divided into 10 sub-categories on the basis of occur-
rence frequency. Coxhead and Nation (2001) stated that because of increased
academic text coverage, in both textbooks and research articles, the AWL list has
been claimed to be a convenient learning tool for L2 learners in academic study.
The AWL has a large impact upon academic writing, vocabulary instruction, and
testing (Simpson-Vlach and Ellis 2010). Moreover, most words in the list are Latini-
ate, which is useful to language learners in general, though it might be easier
for learners with a Romance language as their L1 in particular to remember them
(Coxhead and Byrd 2007).

Some issues with construction of the AWL, however, have been found. The
central issue is linked to the usage and meaning of words (Vongpumivitch et al.
2009). Similarly, the AWL has focused on an EAP context, and thus its contribu-
tion has been limited in terms of the repertoire of the terms relating to learner
occupation or study field. Martinez and Schmitt (2012) have argued that an
academic word list needs not to be created solely from frequently used words
since frequency alone leads to an overabundance of items with an undifferen-
tiated value and “does not necessarily imply either psycholinguistically salient
sequence or pedagogical relevance; common sequences of common words, such
as ‘and of the,’ are expected to occur frequently” (Simpson-Vlach and Ellis, 2010:
490). Meanwhile, Nation (2001) stated that words featured in English academic
writing fall into four categories, which are: frequently used words, academic
words, technical words, and low-frequency words. Given that technical words
occur frequently in particular subject areas, but are uncommon elsewhere, learners may not always be familiar with technical words from outside their own field (Thurstun and Candlin 1998; Nation 2001). Additionally, Nation (2001) asserted that low-frequency words are not often used in some corpora; however, they might be the largest group of words in any field and can include proper names and technical words in other subject areas.

Given such criticisms, researchers have doubted the pedagogic usefulness of academic word lists in the literature. Recent research has tried to combine frequency and new methods to identify specific words in the academic discourse. For instance, Simpson-Vlach and Ellis (2010) created an Academic Formulas List (AFL) through the combination of measures of mutual information (MI) and frequency to investigate the target corpora of academic discourse: MICASE, BNC, Hyland’s (2004) research article corpus, and selected BNC files. From a qualitative perspective, twenty experienced instructors were asked to rate the formulae to discover if the phrases discovered were an expression, a formulaic expression, or a phrase. A correlation analysis was carried out using quantitative statistical methods and qualitative judgement data to ensure the validity and reliability of the instructor insights. Liu (2012) produced a list of the 228 most used multiword constructions (MWCs), which covered various fixed or semi-fixed expressions. The list was generally meant for academic writing across the sub-corpora academic divisions in the Corpus of Contemporary American English and the British National Corpus (BNC). Every MWC identified that words ending with articles (a/the) or another incomplete NP (e.g. one of the) are represented with the ending “det+NP”. Brezina and Gablasova (2015) argued that the word lists in the literature were compiled using different approaches and differ in corpora size. More importantly, they might not reflect the current language use. Brezina and Gablasova, therefore, developed the New General Service List (new-GSL) from four corpora: LOB, BNC, BE06, and EnTenTen12, using a purely quantitative approach and a lemma principle. Based on the average reduced frequency considered from frequency, dispersion, and distribution of the top 3,000 words among the four corpora, Brezina and Gablasova initially generated a stable vocabulary core of 2,122 items. With the aim to create a list of current words, these generated lexical items were then combined with new items frequently occurring in the corpora of BE06 and EnTenTen12, which represent current language use. The finalized list consists of 2,494 items as the lexical core (2,116 base words and 378 current vocabulary). The study also provides evidence of changes in general vocabulary in the English language.

Regarding the collocation functions, Durrant (2009) produced a list of 1,000 academic collocations from the written texts of five faculties: Life Sciences, Science and Engineering, Social-Psychological, Social-Administrative, and Arts
and Humanities. Via the use of WordSmith Tools (Scott, 2004), the collocations were compared with their total frequencies in the 85-million word BNC corpus. Such collocations had to reach a minimum mutual information score of four in all five subject groupings. Martinez and Schmitt (2012) combined the qualitative criteria and frequency in choosing phrasal expressions and individual words. The BNC corpus was chosen to be the corpus source, and WordSmith Tools was employed to search for any 2–4-word strings which were repeated in the corpus a minimum of five times. Additionally, a series of “Auxiliary criteria” and “Core criteria” (Martinez and Schmitt 2012: 308–310) were accounted for to assist in the justification of intuitions in terms of what might be formulaic when choosing multiword expressions for inclusion in the list. Ultimately, a random sampling technique was used to search the derived multiword lexical items line-by-line to see if they had phraseological polysemy. Finally, the PHRASE List consisted of 505 multiword items, claiming to be “useful for pedagogic materials including more multiword items, such as textbooks, graded readers, and language tests” (Martinez and Schmitt 2012: 316). Yet, the benefits of using the PHRASE List remain questionable, also in the studies of Durrant (2009) and Simpson-Vlach and Ellis (2010), primarily since the functions of the multiword items are not given, which can be seen as difficult at first, especially for learners of lower proficiency. This demonstrates that qualitative investigation and functional patterns, as well as quantitative information, need to be considered in constructing academic word lists.

**Multiword combinations**

Because of the variety of formulaic language, scholars have variously defined and used different terms in phraseology research. For example, Altenberg (1998) used the term “recurrent word-combinations” when investigating word patterns which verbally occur in English. The term “lexical bundles” has been used in several research papers (e.g. Biber et al. 1999; Biber and Barbieri 2007; Biber et al. 2004; Conrad and Biber 2004; Chen and Baker 2010; Hyland, 2008a, Hyland 2008b). Schmitt (2004) more often used the term “formulaic sequences”, while “phraseology” and “phraseological patterns” have been used by Charles (2006) and Granger and Meunier (2008) to refer to sets of recurring word combinations. Additionally, the terms “lexical clusters” (Hyland 2008a), “phrasicon” (De Cock et al. 1998), and “n-grams” (Stubbs 2007) refer to multiword sequences. Amongst these terms, Erman and Warren (2000: 31) stated that multiword combinations denote “combinations of at least two words favoured by native speakers in preference to an alternative combination which could have been equivalent
had there been no conventionalization". This is similar to the definition given by Biber et al. (1999) who intuited that they can be fixed expressions or idiomatic phrases, which have a fixed meaning and are understood by language speakers, but cannot be included because lexical bundles are distinct and semantically transparent. Based only upon frequency and distribution criteria gathered from computer programmes, identifying phraseological patterns is a quantitative activity (Biber 2006). Similar to Biber (2006), Cortes (2004) stated that basic techniques used in identifying lexical bundles are word frequency counts, whilst concordance lines, lexico-grammatical profiles, and keyword analysis are used with multiword combinations after they have been identified to pinpoint their functions in context and where they occur in the text. Wray (2002) also discussed that “formulaic sequences” as multiword units which are stored and retrieved from memory as lexical units have become increasingly important for language teachers, researchers, and testers to understand. Likewise, learning and utilizing formulaic language may assist language learners at different levels of proficiency to build fluency and automaticity (Ding 2007; Wood 2006).

Research on phraseology has increased in popularity with its focus on language teaching and learning (e.g. Appel and Wood 2016; Cortes 2004, Cortes 2006; Li and Schmitt 2009; Peters and Pauwels 2015). The research demonstrated that some multiword units or lexical bundles occur frequently in research article corpora. Meanwhile, the investigation of lexical items in the work of students has drawn researchers’ attention to the differences in phraseological patterns between L1 and L2 (Bychkovska and Lee 2017; Pan et al. 2016; Ruan 2016) and between professional and novice writers (Cortes 2006; Peters and Pauwels 2015). For instance, Cortes (2004) compared the function and frequency of lexical clusters in the writings of professional authors and students writing in biology and history. The study confirmed that acquiring and using lexical bundles does not appear to be a natural process. This corroborates Jones and Haywood’s (2004) work which shows that, subsequent to a 10-week instruction period focussing on producing lexical bundles, university students discovered that knowledge of multiword combinations may be technically helpful to express complex ideas, to structure the various writing stages and to attain the required level of formality. Pan et al.’s (2016) study revealed that L2 written texts contain a greater number of lexical bundle types than L1 texts. The structural patterns of the lexical bundles found in these texts are also distinctive.

Additionally, as demonstrated in Li and Schmitt’s (2009) study, the development of students’ repertoires of formulaic sequences over the course is relatively slow, even though these students have majored in language. Although the holistic storage of formulaic sequences has caused controversy (e.g. Siyanova-Chanturia 2015; Durrant and Siyanova-Chanturia 2015), some of the phraseological
research (Durrant 2017; Liu 2012; Martinez and Schmitt 2012; Simpson-Vlach and Ellis 2010) is more complicated since researchers have attempted to create academic word lists that are used in various registers, such as basic conversation, reading and writing (Nation 2001), university textbooks and academic journals (Coxhead 2000), medical texts (Wang et al. 2008), academic writing across disciplines (Durrant 2014, Durrant 2017; Liu 2012), and engineering (Watson Todd 2017). These researchers state that each of the subjects has its own arguments, preferred forms, meanings and syntactical patterns (Martinez et al. 2009) and lexical items in the lists may be caused by the shaping of the disciplines, text selection, and “the particular ways of representing experience” (Yang 2015: 30). Such research, however, gives an insight into a variety of ways to create a pedagogically useful list, allowing us to see the importance and application of corpus-based analysis. Accordingly, the various sizes and types of corpora, as well as the different approaches were considered for this investigation.

Thus, to give Thai novice writers and graduate students support in enhancing their opportunities for scholarly publication, particularly in journals in the TCI database, the list and meanings of multiword combinations may give them a head start in beginning academic research writing tasks. Given the significance of discipline-specific vocabulary, the main objective of this study is to create a multiword combination list useful for writing for publication, which might help language fluency production and which especially helps novice writers and graduate students to effectively create and draft their own research articles. It is thought that learning multiword combinations contributes to the enhancement of communicative competence and that it enables writers to gain the particular rhetorical practices of the texts which they are required to produce (Hyland 2008b). To achieve this goal, this study sought the ways in which language is pragmatically expressed in academic articles by identifying multiword combinations and the associated pragmatic functions typically found in Applied Linguistics research articles.

3 Method

Corpus compilation

The study’s corpus was carefully collected from 120 research articles published in nine journals indexed in the TCI database, in which Thai graduate students and researchers are encouraged to publish their research work. Based on the annual Thai Journal Impact Factors (T-JIF) and the results of journal quality evaluation of the database, all the journals classified in tier 1, which are further
included in the ASEAN Citation Index or ACI database (Svasti and Asavisanu 2007), were chosen. To control any changes in the discipline over time and to enhance the coherence and validity of results, journal samples were restricted to the years 2013–2016. With regard to corpus size, Bowker and Pearson (2002: 45) highlight that “there are no hard and fast rules that can be followed to determine the ideal size of a corpus”. Thus, 120 Applied Linguistics research articles are appropriate in terms of corpus size, since it is manageable and suitable for the study’s objectives and analysis and much useful data and in-depth information can be gained from it. Some factors, e.g. the style of writing and the peer review and copy-editing processes, were not taken into account for the present study. The study focuses on investigating the four internal sections (introduction, methods, results, discussion or IMRD) of the articles; other article sections, were not analyzed in the study, including all the tables, figures, notes, abstracts, references, and appendices in each of the texts. These systematic procedures for corpus compilation yielded approximately 429,438 running words representing the language used in research articles in the discipline of Applied Linguistics. Again, although the entirety of this specialized corpus may appear relatively small in size, compared to previous studies in the literature, we argue that smaller corpora, as specialized ones, are more suitable than large multi-million corpora to identify the connections between linguistic patterning and specialized contexts of language use (Koester 2010). To this end, we were able to gather in-depth information through quantitative and qualitative methods, especially the occurrence of frequent patterns and linguistic items in context.

**Data processing and measures for word selection**

For the investigation of frequency statistics for word sequences in the corpus, n-grams were generated using SketchEngine (SkE) software (Kilgarriff et al. 2004). We first cleaned all texts by removing non-textual content. The edited files were then saved, corresponding to the IMRD sections. Initially, the word list option was used to investigate two-, three- and four-word n-grams, which are referred to here as high frequency formulaic expressions in the corpus. Consideration needs to be given to several issues when identifying multiword units based only on frequency occurrence. First, since n-grams are defined by their occurrence frequency, the frequency cut-offs are arbitrary (Hyland 2008b). The frequency threshold was set; each of the reported frequent n-grams occurred a minimum of eight times in the entire corpus. This cut-off point is determined by the total word number and by the aims of this research to examine the multiword combination usage in the corpus. Second, to compare the n-grams across the article sections, Biber et al.
Attapol Khamkhien and Sue Wharton

Attapol Khamkhien and Sue Wharton (1999) suggested a formula for normalizing frequencies. Based on the length and number of words, the choice of norming to 1,000 words was used in the present study. In regard to the range criteria, we carefully checked all of the generated n-grams to ensure they occurred in at least five files in the corpus, representing the frequency occurrence of such n-grams in at least five articles. This was necessary to guard against subjectivity and idiosyncratic expressions used by individual writers.

It is claimed in the literature that four-word bundles are more phrasal in nature (Biber and Barbieri 2007; Biber et al. 2004; Chen and Baker 2010; Cortes 2004, Cortes 2006; Grabowski 2015; Hyland 2008a, Hyland 2008b). In the analysis, two-word n-gram lists were generated, and we discovered that they mostly appear grammatically incomplete, so that they cannot be understood without the use of nouns or noun phrases (e.g. of the, in the, to the). Simpson-Vlach and Ellis (2010: 493) state that the incomplete bundles are “neither terribly functional nor pedagogically compelling”. Meanwhile, most four-word n-grams (e.g. simple past tense form, intrinsic motivation of English) were found to be content-based lexical items relating to particular subject matter, reflecting an artefact of the writing content. Regarding teachability, they may not have many implications for the entire context and the register in which they are written, in comparison with the n-grams which are grammatically and pragmatically complete units. The three-word n-grams in the analysis seem to be of greater interest than the others as they constitute complete syntactic units as independent meaningful phrases, including some grammatical items, expressing semantic relations (e.g. in order to, as well as), which are not content-based items. Even though their majority does not represent complete structural units (e.g. the use of, the results of), they remain “important building blocks in discourse” (Biber and Barbieri 2007: 270).

As a part of the qualitative process, we extracted content-based strings or noun groups (e.g. language learning strategies, teaching and learning) from the list as they might be useful since they are reflective of the topic or content about which the author is writing (see Appendix). Applying this qualitative criterion, we arrived at a list of 476 potential n-grams, which is quite long for pedagogic purposes. We then applied further selection criteria by progressing through the list item-by-item using a concordancer, searching for “plausibly formulaic” multiword strings (Wray 2009: 41) which realize pragmatic functions or meanings. To ensure high reliability, utility and teachability of the list, five English instructors experienced in EAP, who have publications in peer-reviewed journals, were invited to choose the items which appear to be pedagogically useful for article reading and writing. Specifically, they were invited to rate all the phraseological patterns where, in their opinion, it was worth to learn and teach the multiword
combinations with an eye to research publication writing. Each potential three-word n-gram was chosen by a minimum of three instructors and was included in the final list. The chosen multiword combinations were explored to investigate how they are used by article writers and how they are semantically used in a contextual environment.

4 Findings

Along with using frequency and intuition-based judgements from five EAP instructors, we first generated a list of content word items for anyone interested in how complex noun groups are used in Applied Linguistics articles (see Appendix). Second, a list of 289 functional lexical strings, which appear to be pedagogically useful, was created. The 289-items list is more easily to manage for pedagogic purposes than the 476-items list, but for effective teaching, it is also vital to explore dimensions in the grouping of the target items. The words were then further investigated and categorized according to functional type by looking at them in context and consulting concordance lines. To help the analysis, Biber et al.’s (2004), Hyland’s (2008b) and Durrent’s (2015) functional classifications were used as a guide. All 289 multiword combinations are grouped into four functional categories – research-oriented, text-oriented, stance-oriented and engagement and other functions. However, it should be noted that this list is not intended to be a definitive interpretation of the functional types of multiword combinations, as several of them are found to have multiple functions since they appeared in several sections and contextual environments. Yet, these functions indicate the most salient function fulfilled in an academic context, particularly in the writing of research articles.

Research-oriented functions

Location, procedure, quantification, description, intangible framing attributes

Location

at the beginning
At this stage
from this study
in a text
in the study
in the target
in this group
In this study
the beginning of
the current study
the present study
this study is
Procedure
an analysis of an investigation of analysis of the are expected to as a means by means of can be used data were analyzed Data were collected in order to interview questions were investigation of the is obtained for is used as is used to of data collection of each interviewee the data were the participants in the participants were the process of the questionnaire was the students were The subjects were the use of to answer the to be a to complete the to determine the to find out to identify the to investigate the to participate in to retain the to use a to use the use of the used as a used in the used to analyze used to determine was carried out

Quantification
a corpus of a lot of a number of a part of A total of a variety of all of the as a part each of the frequency of the is one of large number of majority of the most of the number of the one of the out of the some of the the degree of the frequency of the level of the majority of the number of the percentage of the proportion of the scores of

Description
criteria based on meaning of the participants in the the meaning of the the pattern of the study of

Intangible framing attributes
a sense of an important role aware of the development of the good level of knowledge of the mean value of pattern of the reliability of the schematic knowledge of the acquisition of the characteristics of the concept of the development of the effectiveness of the effects of the form of the importance of the influence of the kind of the medium of the nature of the role of the strategy of
the success of
the type of
this statement is
understanding of the
validity of the

*Text-oriented functions*

**Structuring signals, transitional signals, resultative signals, framing signals**

**Structuring signal**
- above table showed
- according to the
- are presented in
- are shown in
- As can be
- As shown in
- based on the
- be seen from

- be seen in
- below illustrates the
- focus on the
- focused on the
- found in the
- illustrates the results
- in the following
- presented in Table

- seen from the
- seen in Table
- shown in Table
- The above table
- the case of
- was based on

**Transition signal**
- as a result
- as well as
- In addition to

- in agreement with
- in line with
- In other words

- On the other
- On the whole
- such as the

**Resultative signal**
- a result of
- agreed that the
- be able to
- be seen that
- by the participants
- consistent with that
- data from the
- differed from the
- employed by the
- finding is also
- findings of the
- findings show that
- followed by the
- found that the
- found to be
- found to exist
- given by the

- have shown that
- indicate that the
- is consistent with
- is similar to
- it was found
- not be able
- of the findings
- of the participants
- of the questionnaire
- of the respondents
- of the student
- of this study
- participants were able
- point out that
- result shows that
- results from the
- results of the

- revealed that the
- show that the
- stated that the
- study demonstrated that
- study found that
- suggests that the
- the data from
- the findings from
- the findings of
- the participants had
- the respondents had
- the result of
- The results from
- The results show
- This finding is
- This indicates that
- This means that
This suggests that was found that were reported to
to the participants was found to were consistent with
used by a

**Framing signal**
As for the in the process the context of
for further research in the use the other hand
from the context part of the the part of
in relation to purpose of the the purpose of
in terms of terms of the with regard to
in the future the basis of

**Stance-oriented and engagement functions**

**Stance feature**
are likely to highly related to need to be
are more likely is important to related to the
be aware of is possible that seems to be
be concluded that is suggested that should be conducted
be related to it can be should be noted
be said that it could be similar to that
because of the It is also so that they
can also be it is necessary students should be
can be seen it is possible there is a
compared to the It must be there is no
can be seen compared to those It should be This can be
can be seen considered as a it would be This is because
can be seen considered to be likely to be to be able
can be seen contribute to the might not be to be aware
can be seen due to the more likely to was able to

**Engagement feature**
be noted that

**Other functions**
developed by the is not only This is in
exist at the research has been was divided into
fact that the study aims to was employed to
identified according to the fact that was identified according
is in line the sense that was used to
The word combination functional analysis, from a pedagogical viewpoint, is essential to understanding the value of the combinations as teaching items. However, language utterances can vary widely according to use, interpretation, socio-cultural factors, social conventions, etc. Since there is no context-free correspondence between structural patterns and pragmatic functions, we argued that each of the phraseological units included in the list can therefore express more than one pragmatic function. Consequently, we concentrated on a small selection of lexical phrases, identifying their dependency on context and topic by using concordance lines to examine how these words are used in the text in terms of salient pragmatic functions. The investigation results and their descriptions are as follows:

**Research-oriented functions** help the writers with the structure of their research activities and experiences. This group is the largest category including those which refer to research location or place, procedure, quantification, description and topic of the research, and intangible framing attributes.

1. **Analysis of variance (ANOVA) was used to determine the comparability of groups at the beginning of the study.** [M 26]

2. **A corpus-driven approach was thereby applied to an analysis of Jane Austen’s six major novels in order to see how well this method works with literary texts.** [M 26]

3. **The most cited strategy used is rereading (Table 2, #17) which is a very basic and traditional strategy although some of the participants stated that they reread with a purpose, they only reread and focused on the important part.** [D 25]

4. **Moreover, this result supported the study of Kim and Petraki (2009, p. 72) stating that the recognition of L1 importance declined from the advanced group, and increased in the intermediate, and the beginning respectively.** [R 26]

5. **To investigate possible ways to encourage the development of Thai learners’ speaking skills, this study aims to research their attitudes and motivation in learning to speak English.** [I 1]

As can be seen, in (1) *the beginning of* is an example of the location sub-category referring to location or spatial reference points in the text. The cluster *in order to*
as in (2) is classified in the procedure sub-category, which indicates the objective of the approach used for analysis in the study. Quantification features (some of the) as in (3) refer to participant quantity. This group relates mostly to the number of samples or participants involved in research activities, data, researchers, and related research. The string (the study of) in (4) in the description sub-category describes the research’s physical properties that the writer compares the research findings with. Intangible framing attributes (the development of) as in (5) refer to learner abstract properties, such as speaking abilities and development.

**Text-oriented functions** deal with text meaning and its organization. Transition signals, structuring signals, resultative signals and framing signals of the text are included in this category.

(6) *An ability to hold a conversation during flights in English is just as important as listening skills as well as service functions in the role of cabin crew.* [D 1]

(7) *The lists of keywords, semantic fields and grammatical categories in JA are presented in* Tables 1–3 below, starting from the item with the highest degree of keyness. [R 26]

(8) *The interview revealed that the high vocabulary subjects seemed to have positive attitudes towards English while the low vocabulary subjects seemed to have negative attitudes towards the language.* [D 26]

(9) *The purpose of this study was to determine how typically developing children and children with autism construe their experience of the world around and inside them in producing their narratives.* [I 26]

In (6), structuring signals (are presented in) denote parts of the text, which helps to direct the reader to visuals and/or particular sections of the text. In (7), transition signals (as well as) indicate text structure, which directs readers to the information’s location (in Table 1–3). This sub-category includes phrases showing the relationships of addition, contrast, or equivalence between elements, called discourse markers in Biber et al.’s (2004) classification. Resultative signals refer to causative or inferential relationships between elements. The string the purpose of in (9) is used to state the study objective, showing what research was conducted.

**Stance-oriented and engagement functions** express epistemic judgements, attitudes, evaluations, and degrees of commitment regarding the claims which are being made. The findings for this category corroborate Simpson-Vlach and Ellis’ (2010) statement that the formulae are associated with knowledge claims, expression of certainty or uncertainty, beliefs, thoughts, or claims made by others. Hyland (2008a, 2008b) and Biber et al. (2004) state that stance-oriented functions also express a degree of migration, tentativeness and claim possibilities.
This function category includes two functional subgroups – stance-oriented and engagement functions. Yet, here only one item is included in the engagement functions.

(10) *English language learners are likely to use the language with people from various language and cultural backgrounds.*  

(11) *It should be noted that even though the respondents strongly aspired for native-like pronunciation, they were aware that native-like pronunciation is not the only requirement for successful communication.*

Stance-oriented functions, such as *are likely to* in (10), reflect the writing’s evaluative nature. With this expression, the writer expresses his or her interpretations and attitudes towards statements in terms of possibilities. In (11), the string *be noted that* as grouped in the engagement functions indicates the statement’s importance. The writer would like to incorporate the active role of potential readers. In this context, Hyland (2001: 552) points out that the exchange between writer and reader is established when readers are considered as “real players in the discourse rather than merely as implied observers of the discussion”.

**Other functions** refer to the meanings which vary widely depending on the particular context: interpretation, socio-cultural factors, social conventions, etc.

(12) *This finding is in line with Sarani and Kafipour (2008), who reported that L2 learners did not use dictionaries appropriately.*

(13) *These samples were divided into two groups – 20 good readers and 20 poor readers – based on their grades in 4 previous reading courses.*

The n-gram *is in line (with)* in (12) is commonly used when writers compare their research results with previous research. Subsequently, in (13), the string *were divided into* describes the study’s participants regarding the research method used in an experiment. Essentially, the multiword combinations’ defined functions and meanings included in this category are dependent upon the possible environmental contexts in which they are used.

## 5 Discussion and conclusion

This study examined Applied Linguistics research articles by using repeated frequent three-word sequences and psychological judgements. The aim of the research was to create a pedagogically useful list of multiword items and to provide
their semantic and pragmatic functions to aid the task of research manuscript writing for publication. Based on a corpus-based, qualitative approach and the opinions of five EAP instructors, a list of 289 three-word multiword combinations for teaching research article writing in English was generated. We assessed how much phraseology contributes to article writing by investigating lexical cluster pragmatic functions included in the list. A combination of qualitative and quantitative approaches in list development has its advantages. The combination of objective and subjective criteria is seen as a complementary perspective, whereas quantitative analysis being qualitatively validated is also crucial, offering a powerful way to understand texts. Meanwhile, the inclusion of instructor insights is seen as another selection criterion which can maximize the pedagogical usefulness of the list. Taking all of these aspects together – quantitative frequency, qualitative judgements about what are meaningful phrases, and inputs from experts in the field in considering those useful phrases – demonstrates a thorough perspective on textual analysis to receive specific and in-depth information. These methodological choices ensure that the word list is developed in a transparent and reliable way, contains items which are frequently used, and can potentially be useful. As for the top-down perspective, the quantitative approach, like the corpus-based investigation, shows that the greatest range of content words is in the article corpus but is not included in the final list, since the list’s pedagogic purpose is also a principal objective. In the qualitative approach, based on context dependency, some multiword sequences seem to possess multi-functionality, appearing across several sections. For example, ‘according to the’ could appear in the Methods, Results and Discussion sections, while ‘some of them’ is found in every section across the text. This bottom-up perspective identifies the functional types in terms of context and occurrence in the text. The finding concurs with Simpson-Vlach and Ellis’s (2010) research that currency and frequency alone cannot assure functional utility, rendering teachability and pedagogic value. In this regard, the list of content-word strings is useful and has meaning for researchers who are interested in the use of the English language in research literature and in and how complex noun phrases are used in the publication of articles. We also argued that semantic and pragmatic criteria are more meaningful than those based on frequency, and this combination of research methods is, therefore, substantially important in developing a list of functional strings.

Since the phraseological units selected for the present study are syntactically complete units, their characteristics are distinct from those of lexical bundles in previous research (Biber et al. 2004; Cortes 2004, Cortes 2007, Cortes 2013). One of the potential reasons for this is that Biber et al.’s (2004) and Hyland’s (2008a, 2008b) classifications are derived from the analysis of a huge corpus, including various disciplines and registers. Meanwhile, Coxhead’s (2000) and Gardner and
Davies’s works (2014) focused on word families in the academic lexis. In addition, the current study focuses exclusively on three-word phraseological patterns, rather than four- or more word bundles (e.g. Biber et al. 2004; Conrad and Biber 2004; Cortes 2004; Hyland 2008a, Hyland 2008b). The pragmatic functions of the lexical items found in the context are distinctive, reflecting the topic-specific and language use in Applied Linguistics research articles. Given that the aim of this research was to help novice writers and students to draft their research manuscripts effectively, examining a specific corpus from a single discipline is likely to be beneficial since it yields more specific functional characteristics (Durrant 2017).

As far as the pedagogical purpose is concerned, novice writers and students should be aware of types of lexical items and their relation to information structure and/or discourse function. Csomay (2013) suggested that students often don’t consider that multiword items and grammatical patterns can indicate a change in text type within discourse. The list and pragmatic functions suggested in this study can serve as the basis for proficient academic writing. Thus, when designing an academic writing course for publication, instructors could make full use of the list and integrate a description of this study into their instruction. Hyland (2008b) suggests that writers are expected to stick to the linguistic rules of language and comply with the intended readers’ expectations via the implementation of potential lexical clusters of the discourse in question. Students and novice article writers should therefore have the required knowledge about the use and pragmatic functions of multiword combinations applied in a given section when preparing their research manuscript. Instructors might use the knowledge and multiword items list of this study by implementing some activities which feature different lexical cluster types, with an emphasis on fostering the expressive skills of their students and on how to use the clusters for communicative purposes. Moreover, to improve the usefulness of the list, instructors may describe patterns of use or structural “frames” rather than solely teaching the multiword combinations. For example, ‘to answer the’ is probably going to have ‘first, second, third question or research question’ as the next element and to indicate this would make the list much more useful. This may develop the students’ proficiency and experience in using contextually appropriate words while writing academic texts (Pan et al. 2016). Instructors might also draw their students’ attention to the words in the list and encourage their use in assignment writing. This supports Wood’s (2015) notion which suggests that formulaic sequence knowledge can be used with sensitivity. For example, formulaic sequences can be integrated into language pedagogy by using them with form-focused lesson, instruction, and specific types of activities, such as searching corpora for concordances of sequences, or replacing single words with sequences. To introduce different lexical clusters
which act as various pragmatic functions and to raise student awareness about the importance of this language phenomenon in academic contexts will help students in drafting articles which meet the required levels current in academic and/or research communities (Coxhead and Byrd 2007; Martinez and Schmitt 2012).

At a more advanced level, the comprehensive approach to selecting pedagogically useful phraseological patterns included in the current study is a starting point for setting vocabulary goals for advanced language courses, especially in terms of guiding graduate students in independent study. Hyland and Tse (2009) suggested that a good method to prepare students for studying is not to search for universally appropriate teaching items. However, regarding a genre-based approach to teaching, we would argue that instructors can take advantage of selecting phraseological units to create academic word lists which fit a specific classroom setting, context and pedagogic purpose. They could explain to the students that the selected phraseological patterns are some of the important linguistic features they might encounter in academic settings and especially in engaging in writing research articles. In particular, the selection criteria can help instructors in selecting texts and developing learning-related activities to promote student sensitivity to the importance of lexico-grammatical features and phraseological patterns which occur frequently in the text.

However, caution is required in applying the findings and the lists to pedagogy as the corpus of this study stems from a single discipline. The results and the list should be considered as only illustrative and have restrictions because they relate only to Applied Linguistics research articles published in English which are indexed in the TCI database, rather than to English articles in various other disciplines which might not be included in the TCI. Regarding the selection criteria used in this study, the items chosen for inclusion in the pedagogically useful list of functional n-grams and content-based strings are not supposed to be representative of the entirety of all phraseology used in articles published in this field in English. There might be phraseological patterns, pragmatic functions and complex noun groups which are not present in the current study. Yet, the study’s findings are considered meaningful enough for those intending to publish their research in journals, especially the ones included in the TCI database. Other multiword combinations might not be conclusive and are not included in the pedagogically useful list. Additionally, the methodology and scope of this study should be considered. Firstly, the number of articles analyzed in this study is relatively small and specific. To generalize the findings, a bigger corpus size might achieve an improved yield and represent a better global picture of multiword combinations used in the articles. Secondly, it was found in the corpus that 3-grams are more useful than bigrams and longer grams. However, it should be taken into account that 3-grams generated from the corpus in this study might be the effect of the corpus size. We acknowledge that longer sequences
might be useful as they can help reveal the semantic and pragmatic functions from the context in which the strings are used. Therefore, the bigger the corpus, the more interesting expressions can be revealed by a longer n-gram calculation. It is also acknowledged that the list is only raw material that will need further work to prove how it is useful to EAP writers. A process could be envisaged by which the article writers are given the list as well as the database, and they can then search for the specific context using a concordancer. An alternative approach would be to take other eminent statistic criteria such as MI-score and formula teaching worth (Simpson-Vlach & Ellis 2010), together with the careful selection from EAP instructors, to support an identification task for multiword units useful for pedagogic purposes and to obtain a more refined pedagogically useful list (Salazar 2011). Additionally, it is crucial to encourage and teach students to consult other reliable resources when encountering multiword items and experiencing difficulty in reading and writing. Despite the scope for future research, the descriptive results here remain crucial for EAP instructors in developing instructional materials in teaching writing for scholarly publication. This might help graduate students and novice writers with the preparation of manuscripts for publication.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments and suggestions.

References


Cortes, Viviana. 2006. Teaching lexical bundles in the disciplines: An example from a writing intensive history class. Linguistics and Education 17(4). 391–406.

Cortes, Viviana. 2013. The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. Journal of English for Academic Purposes 12(1). 33–43.


Appendix

List of content-based n-grams generated from the corpus of 120 research articles

vocabulary learning strategy/strategies  multiple choice options
foreign language anxiety  general English proficiency
English language learning  high proficiency group
child/children with autism  low listening ability
Jane Austen’s novels  paper-pencil peer feedback
typically developing children  positive politeness strategies
reading for pleasure  codes corrective feedback
foreign language classroom  language learning strategies
English language teaching  negative politeness strategies
consonant segmental phonemes  frequently used strategies
listening ability group  perceived communication mobility
written corrective feedback  no significant difference
low vocabulary subjects  students’ writing ability
low proficiency group  teaching cultural content
language classroom anxiety  vocabulary size test
high vocabulary subjects  English major students
newly learned words  Information gap tasks
academic listening comprehension  overall mean score
high listening ability  proves writing approach
significant linguistic features  vocabulary learning problems