THE JOURNAL OF ASIA TEFL Vol.10, No. 4, pp. 71-96, Winter 2013

Analyzing the Journal Corpus Data on English Expressions Across Disciplines

Sayako Maswana
Waseda University, Japan

Toshiyuki Kanamaru
Kyoto University, Japan

Akira Tajino
Kyoto University, Japan

While a number of English for Academic Purposes (EAP) studies have examined the linguistic features extracted from research paper corpora based on “frequency” and “range,” few attempts have been made to examine whether these criteria meet the perspectives of actual readers and writers of research papers, or to explore the relationship between linguistic features and textual organization. This paper attempts to reveal the nature of expressions researchers consider “useful” by comparing them with those generated based on frequency and examining their relationship with structural segments (*moves*) of research papers. In the study, over 20,000 English expressions were extracted from the perspectives of as many as 51 researchers from 15 disciplines at a research-oriented university in Japan. With a focus on one- and four-word expressions, the main findings show use of salient adverbial expressions, noun phrases without post-modifier fragments in the researcher-selected expressions, and the existence of expression-dependent and expression-independent moves. The findings reflect researchers’ multiple perspectives, such as linking and creating words and phrases, which would increase and improve learners’ means of expression. The results of the study provide an interdisciplinary expression corpus with contextual information, which may be used to complement frequency-based expression resources.

Keywords: English expressions, vocabulary, lexical bundles, moves, research papers, collaboration

INTRODUCTION

Writing research papers in English has become a requirement for many students at research-oriented universities in an EFL environment. The mastery of appropriate vocabulary and multiword expressions is considered crucial for successful research paper writing, and research is available that has examined these expressions across different disciplines (e.g., Coxhead, 2000) as well as in one particular discipline (e.g., Martínez, Beck, & Panza, 2009). Coxhead (2000) provides 570 words common to research papers across a broad range of disciplines, while Martínez et al. (2009) investigate discipline-specific vocabulary.

Similarly, multiword expressions have been investigated for general academic writing (e.g., Biber & Conrad, 1999; Martinez & Schmitt, 2012; Simpson-Vlach & Ellis, 2010) and for specific disciplines (e.g., Cortes, 2004; Hyland, 2008). In many studies, multiword expressions are defined as multiword combinations that frequently occur in a wide range of texts and are variously referred to as multiword constructions, lexical bundles, word clusters, and formulaic sequences.

Advancement in the corpus study has made it possible to extract and introduce into teaching research findings based primarily on frequency. While previous studies have provided some insight into the nature of vocabulary and fixed expressions in academic writing, the use and applicability of their results are somewhat limited in classroom settings. Some reasons might include that the corpus used was limited in its range of disciplines, that it lacked contextual information, or that the frequency-based extraction was not sufficient for providing a useful list of expressions for academic writing.

To overcome these challenges, 51 researchers from 15 disciplines at a research-oriented university in Japan participated as analysts in a study to select expressions[1]. The study included tagging the code of textual organizational information of the research papers. The research question addressed in the study is: What kind of expressions do researchers from various disciplines find useful for writing academic papers? The study examined the nature of the expressions by focusing on one- and four-word expressions in their forms and relationship to the textual organization and comparing them with frequency-based extraction. The study aims to provide a disciplinary researcher's perspective of useful expressions and the results of the study are expected to provide a practical resource for academic writing instruction, particularly for undergraduate and graduate students at universities in an EFL environment who have to write research papers in English.

One- and Four-word Expressions in Research Papers

Among the various types of expression used in research papers, vocabulary is one of the most studied areas. Many studies have attempted to compile a vocabulary list common to a wide range of disciplines and specific to one discipline (e.g., Coxhead, 2000; Farrell, 1990; Tajino & Kanamaru, 2011; Xue & Nation, 1984). These studies generally agree that, in academic papers, general vocabulary common to any type of text, often represented by the General Service List (GSL), accounts for approximately 80% of most academic texts; academic vocabulary common to many disciplines accounts for 8–10%; and technical vocabulary particular to each discipline accounts for 5% (Nation, 2001). This has direct implications for teaching academic

[1] The term "expression" is adopted in this paper because expressions include one-word and multiword constructions, phrases, clusters, chunks, and bundles; however, expressions are not necessarily fixed or semi-fixed.

writing because general vocabulary is supposed to be learned before entering university. Furthermore, academic vocabulary is learned at the beginning level (first and second years) of university, while technical vocabulary is learned in disciplinary courses taught by content instructors. Chung and Nation (2003) examined technical vocabulary by having disciplinary experts scale for each word according to the level of relatedness to the disciplinary content. In comparison, the specialist's rating scale for identifying technical vocabulary was the most accurate against the other three approaches, that is, derived from the clue in the text, dictionary-based, and computer-generated (Chung & Nation, 2004). Studies concluded, however, that computer extraction has been the most realistic and relatively reliable method thus far. Chung and Nation's study (2004) is important in its attempt to investigate the underlying assumption that automatic retrieval based on frequency and range reflects the vocabulary needed to express disciplinary content as experts.

Research to evaluate academic vocabulary lists is also available, such as the study of Hyland and Tse (2007), which demonstrated that disciplinary variation is significant in the area of so-called academic vocabulary. The question still remains: to what extent the list generated based on frequency in a corpus is actually helpful for learners? There are ongoing efforts to retrieve vocabulary that would best help students write their academic papers by using various corpora and different retrieval methods and selection criteria.

In addition to vocabulary, academic texts have been said to contain distinct multiword expressions (Biber & Barbieri, 2007; Biber & Conrad, 1999; Biber, Conrad, & Cortes, 2004; Biber, Johansson, Leech, Conrad, & Finegan, 1999; Cortes, 2004, 2013). *Lexical bundles*, as they are termed by Biber and his colleagues, are recurrent combinations of words that are retrieved based on frequency and dispersion criteria from selected corpora (Biber et al., 1999, pp.990-993). Among the multiword lexical bundles, the four-word bundle is the most studied because the size of retrieved expressions is usually manageable for analysis; three-word bundles, many of which are important expressions, are often subsumed into four-word bundles (e.g., *as a result* and *as a result of*, Cortes, 2004, p. 401) (Byrd & Coxhead, 2010; Chen & Baker, 2010; Cortes, 2004).

While these frequency-based vocabulary and four-word bundle studies have offered insights into the use of expressions in academic writing as well as the pedagogical implications, expressions in the previous studies, except Cortes (2013), were presented without contextual information. Because those expressions were retrieved from the corpus, researchers made the assumption that contextual information can be seen any time one searches for the collocation of the expression. Even if one knows the words that come before and after a target expression and its surrounding sentences, this does not provide information regarding the paper's overall textual structure. A research paper corpus containing contextual information is rare due to the difficulty of fully understanding the content of multidisciplinary research papers by linguistic analysts. The broader contextual information provided by disciplinary researchers might offer additional help for learners to be able to associate multiword expressions with the particular content they would like to express. Another issue is that frequency-based expressions—particularly multiword expressions—have low face value because students are already familiar with these expressions at the tertiary level (Byrd & Coxhead, 2010). These issues are considered in the present study.

THE STUDY

Participants

Participants in this study included 51 researchers from 15 faculties and graduate schools. The faculties included Integrated Human Studies[2], Letters, Education, Law, Economics, Science, Medicine, Pharmaceutical Sciences, Engineering, and Agriculture. The graduate schools included Energy Science, Asian and African Area Studies, Informatics, Biostudies, and Global

[2] The faculty consists of inter- and multi-disciplinary fields such as cognitive and neurological sciences.

Environmental Studies. The native status of the researchers was not considered because these researchers are active members of the discourse community who read and write the target texts.

Procedure

This study used the Kyoto University academic paper corpus established in 2008 as a language resource for academic writing education. The corpus was made possible with the help of researchers from all faculties and research institutes at the university. The corpus of approximately 15 million words was compiled with a selection of 2,052 research papers from 170 highly recognized international journals (see Kanamaru, Maswana, Sasao, & Tajino, 2010 for more details).

A session was held by the authors for the purpose of informing 51 researchers about the objectives of the research and the textual analysis procedure. The researchers selected articles for analysis that were close to their own fields of research. Each researcher analyzed approximately 100 pages, reaching a total of 423 papers, or approximately 2.5 million words. The number of articles analyzed is listed in Table 1.

The researchers tagged expressions or language features that they considered useful in writing research papers. Expressions could consist of a word, a phrase, a collocation, or even a sentence. The definition of the term *useful* was intentionally constructed loosely in the analysis instructions, because the study hoped to select as many expressions as possible from the disciplinary researchers' perspectives. Along with the tagging of expressions, the analysts coded textual structure based on Swales' (1990, 2004) move analysis by referring to the classification codes created by the authors. We modified move categories suggested by Kanoksilapatham (2005) and Nwogu (1997) to adapt to a wide range of disciplines involved in the study (see Maswana, 2013 for more details about the move analysis). The researchers completed an open-ended questionnaire after the selection of expressions and coding, which included items asking for their comments on the selection of useful expressions and on their tagging, including where they had difficulties. The survey also asked for comments on the procedure and project.

TABLE 1
Number of Articles Analyzed by Faculty/Graduate School

Faculty/Graduate School	No. of Articles	Faculty/Graduate School	No. of Articles
Faculty of Integrated Human Studies	8	Faculty of Engineering	59
Faculty of Letters	47	Faculty of Agriculture	58
Faculty of Education	13	Graduate School of Energy Science	9
Faculty of Law	4	Graduate School of Asian and African Area Studies	23
Faculty of Economics	8	Graduate School of Informatics[a]	8
Faculty of Science	58	Graduate School of Biostudies	33
Faculty of Medicine	37	Graduate School of Global Environmental Studies	14
Faculty of Pharmaceutical Sciences	44	Total	423

[a] The informatics researcher selected articles from the Faculty of Engineering

Data Analysis

After having collected all the extracted data, we made a corpus of expressions with contextual information. In the course of data analysis, we contacted the researchers when questions concerning the content of the text arose. The expressions were categorized according to the number of words comprised in expressions in order to allow comparative analysis with frequency-based expressions. For single-word expressions, the study did not use the concept of *word family* to count words so that we would better understand the function of the word in a sentence. A word with two separate functions thus counted as two separate words. For example, *indicating*, used in a participial phrase, and *indicates*, used in simple present tense, are counted as two words. We hoped to reflect the intention of the researchers who selected expressions in terms of the word's usage in the text. For comparison, we created a list of the corpus's most frequently used single words for the same number of expressions selected by the researchers, using Wordsmith (version 5.0). We used the same criteria as the researcher-extracted vocabulary to count the words extracted by Wordsmith. We also extracted four-word expressions based on frequency and range for the same number of four-word expressions selected by the researchers.

RESULTS

Our study has extracted 20,193 expression types (34,928 tokens[3]) that vary in the number of words, from one word to one sentence consisting of 30 words, as well as in their function in the text. The list reveals what the participating researchers perceive as "useful" expressions when writing research papers in English. It shows striking differences from frequency-based computed lists of expressions often suggested for learners. Here we will focus on the results of one- and four-word expressions because they are the two types of expressions that have been extensively studied and thus have comparable research to examine for their characteristics.

Single-word Expressions

There are a total of 304 single-word expression types[4] (1,858 tokens) in the selected useful expressions. To examine researcher-selected and frequency-based expressions, we first categorized the words in terms of word level.

[3] This includes the same expression selected by several researchers as well as the same word selected by a researcher several times in one paper.

[4] We did not make a distinction between uppercase and lowercase or with a comma and without a comma.

TABLE 2
Profile of Researcher-extracted and Frequency-based One-word Expressions

Word level	Researcher extracted	Frequency based
General Service List 1–1000 words	108	234
General Service List 1001–2000 words	19	14
Academic Word List words	82	38
Off-list words	95	18
Total	304	304

The two lists share 37 words, as follows: *for, but, also, using, if, however, first, see, because, analysis, then, results, where, thus, specific, based, observed, although, significant, given, while, second, since, therefore, respectively, without, measured, following, further, whether, here, including, rather, whereas, compared, significantly,* and *samples*. Except for the second 1,000 most frequent words of the GSL[5], statistical differences among the frequencies of the words in each category are observable. It is important to note the nature of the differences and the fact that quite a few of the first 1,000 most frequent words of the GSL were selected by the researchers. As many of the researchers possess doctoral degrees, the selection of the basic 1,000 words seems counterintuitive. The first 2,000 words are not usually instructed in academic writing classes.

The off-list category includes proper nouns, such as Ca^{2+}, in the database; technical terms such as *anamnesis*; and combined words such as *up-regulation*. Indeed, the one-word expression list includes 23 words that are combined with more than one hyphen or slash; seven nouns such as *up-regulation* and *diet-induced-obesity*; 14 adjectives such as *double-mutant*

[5] GSL stands for the General Service List (West, 1953), which has selected the 2,000 most common English words.

and **-deficient*[6]*;* one adverb (*self-consciously*); and one conjunction (*and/or*). *So-called* was considered an exception because it is a widely recognized adjective.

Next, we examined the primary function of the selected one-word expressions in the texts. Table 3 shows the part of speech of each of the 304 researcher-selected words[7]. If a word represented multiple parts of speech, the function it served in the particular text was assigned as its part of speech.[8] For example, "*Below*, …" was labeled as an adverb rather than a preposition.

TABLE 3
Distribution of Parts of Speech Among Vocabulary Selected by the Researchers

Verb	Adjective	Noun	Adverb	Conjunction	Abbreviation	Preposition	Total
78	28	45	124	13	3	13	304

The most frequent part of speech was the adverb, at 124 instances, followed by the verb, at 78. It is important, however, to take a closer look at the verbs. Among the 78 verbs that appeared, 15 were in *-ing* forms (e.g., "The improvement in fit between 1987 and 1998 profiles was negligible, *indicating* process LP is no longer important") of which five appeared at the beginning of a sentence (e.g., "*Assuming* the Nash bargaining solution where everyone earns . . ."), causing them to function adverbially. In addition, 10 verbs appeared in the imperative form, also placed at the beginning of the

[6] Proper nouns or numbers were replaced with an asterisk (*) by the researcher as they can be replaced by any number of nouns.

[7] Distribution of parts of speech was counted only for the researcher-selected expressions. This is because the researchers selected expressions deemed useful; therefore, they could select the same expression several times as well as only once. The frequency-based expressions, on the other hand, count the same word every time it appears, disregarding its usefulness in discourse. The same reason goes for the number of expressions in each move presented in Table 4 and Table 7.

[8] If the word serves several functions in a text, the function that occurred most frequently was considered as the word's part of speech.

sentence (e.g., "*Fix* an irreducible subvariety Q ⊂ G (k, n), and . . ."). In this last example, the researcher should know the general meaning of *fix*, yet he selected the word because of its particular meaning and the form (imperative) used in this context. In fact, 1,038 of the total selected one-word expressions (tokens) were positioned at the beginning of a sentence.

Table 4 is a profile of the researcher-selected one-word expressions in terms of their position in regard to a whole article indicated in *move*. One-word expressions are mostly selected where the author describes the results of the research (*move 10*: explaining specific research outcomes in the conclusion section, and *move 7*: reporting results in the results section), followed by *move 2*: reviewing related research in the introduction section, and *move 5*: describing experimental procedures in the methods section.

TABLE 4
Number of One-word Expressions by Move (in token)

Move	No. of Expressions
a: Abstract	52
1: Presenting background information	102
2: Reviewing related research	203
3: Presenting new research conducted by the author(s)	86
4: Identifying the source of data and the method adopted in collecting data	66
5: Describing experimental procedures	162
6: Describing data-analysis procedures	109
7: Reporting results	327
8: Commenting on results	129
9: Highlighting overall results and their significance	30
10: Explaining specific research outcomes	358
11: Stating research conclusions	47
12: Interpreting and analyzing the original text	80
13: Developing the author's interpretation of the original text	23
14: Arguing with critics	1
15: Describing the results of comparative law research	1

16: Presenting the theoretical model to serve as the basis for experiment design	2
17: Describing the procedure to test the prediction of the theoretical model	1
18: Discussing the expected outcomes	15
19: Describing the lemma[a]	7
20: Describing the proposition	7
21: Describing the theorem	15
elsewhere[b]	35
Total	1,858

[a] Lemma here refers to a subsidiary proposition.

[b] Elsewhere includes footnotes and figure and table captions.

Four-word Expressions

Similar to vocabulary research, four-word expressions have been closely examined based on frequency and range (e.g., Chen & Baker, 2010; Hyland, 2008). Our study selected 2,783 four-word expressions (5,763 tokens). We have also extracted 2,783 four-word expressions in terms of frequency and range. In fact, 2,783 four-word bundles are too many to extract from our corpus based on criteria employed in previous studies, such as a minimum of 10 times per million words in five different texts (Biber et al., 1999, p.992). Thus, for the purpose of comparison, the four-word expressions were extracted based firstly on frequency. Next, to avoid idiosyncratic use of one author, the cut-off range was set at two, meaning that the expression should appear in at least more than two texts. The minimum frequency and range of the expressions in the frequency-based list turned out to be eight and two. The two lists share 303 expressions. Following the categorization of Chen and Baker (2010), we created a structure profile for the researcher-selected and frequency-based four-word expressions (see Table 5)[9].

[9] We have added the adjective-based (e.g., *consistent with previous research*), adverb-based (e.g., *finally and most importantly*), and conjunction-based (e.g., *if and only if*) to Chen and Baker (2010).

TABLE 5
Structure of Four-word Expressions

	Researcher selected	Frequency based
Noun-based	258	1,260
Adjective-based	43	46
Adverb-based	37	6
Conjunction-based	41	18
Preposition-based	536	741
Verb-based	1,868	712
Total	2,783	2,783

For the researcher-selected four-word expressions, the verb-based construction is by far the most frequent structure. Contrarily, the noun-based construction, followed by the preposition-based construction, is predominantly used in the frequency-based expressions, which supports the results of previous studies (Biber, Conrad, & Cortes, 2004; Byrd & Coxhead, 2010; Liu, 2012). Some differences exist between the two lists in terms of noun-based constructions. Although a noun phrase with post-modifier fragment (e.g., *a high degree of*) is dominant in both lists, more noun phrases without post-modifier fragments (e.g., *a more recent example*) exist in the researcher-based list than in the frequency-based list (e.g., 19 out of 120 expressions starting with a/an in the researcher-selected list compared to 3 out of 119 in the frequency-based extraction). Indeed, there are quite a few noun phrases without any prepositions in the researcher-selected expressions such as *most dramatically accelerated change*, *an intuitively satisfying way*, and *free radical scavenging properties,* which are rare in the frequency-based extraction.

It is also important to consider that verb-based expressions include participle constructions such as in "We search for the expected decay of adjacent SNP linkage disequilibrium (LD) at recently selected alleles, *eliminating the need for* inferring haplotype," which in practice works adverbially.

TABLE 6
Most Frequent First Words in Four-word Expressions

	Researcher-based	Frequency-based
is, are, was, were	358	202
in	245	217
it	215	49
has, have, had	154	26
as	142	61
a, an	120	119
the	76	413
can, could	71	24
may, might	51	10 (0 might)
one	30	13

Table 6 shows the 10 most frequent initial words in four-word expressions chosen by the researchers and their corresponding number in the frequency-based extraction. Reflecting the dominant verb-based construction, "is, are, was, were" are the most frequent first words in the researcher-based list, which implies the inclusion of many passive-voice expressions. For the frequency-based list, other first words include many prepositions such as "of" for 122 words, "to" for 78, "for" for 62, and "at" for 53 words, which constitute the 10 most frequent first words. It should be noted that certain words and prepositions dominate in the frequency-based extraction (e.g., "the" counts in 15% of the total expressions) and auxiliary verbs (can, could, may, might) are placed among the top 10 for the researcher-selected list while not in the frequency-based list. Although outside the list, the researcher-selected list includes 10 expressions starting with "should." Other frequent expressions outside the list start with "provide…" (21 expressions), "consistent with . . ." (14 expressions), and "let us . . ." (12 expressions) while none of these expressions is included in the frequency-based list. It is also important to note that 2,370 of the total expressions (tokens) came at the beginning of the sentence, which accounts for about half of the total four-word expressions selected by the researchers.

Table 7 is a profile of the researcher-selected four-word expressions in terms of relative position to a whole article. Very similar to the researcher-selected vocabulary, *move 7*: reporting results in the results section and *move 10*: explaining specific research outcomes in the conclusion section are dominant, followed by *move 2*: reviewing related research in the introduction section. The only difference is that instead of *move 5*: describing experimental procedures of the methods section, *move 8*: commenting on results in the results section is the fourth most common location of the four-word expressions.

TABLE 7
Number of Four-word Expressions by Move (in token)

Move	No. of expressions
a: Abstract	286
1: Presenting background information	346
2: Reviewing related research	657
3: Presenting new research conducted by the author(s)	347
4: Identifying the source of data and the method adopted in collecting data	284
5: Describing experimental procedures	293
6: Describing data analysis procedures	335
7: Reporting results	1,002
8: Commenting on results	581
9: Highlighting overall results and their significance	127
10: Explaining specific research outcomes	955
11: Stating research conclusions	163
12: Interpreting and analyzing the original text	139

13: Developing the author's interpretation of the original text	66
14: Arguing with critics	1
15: Describing the results of comparative law research	4
16: Presenting the theoretical model to serve as the basis for experiment design	3
17: Describing the procedure to test the prediction of the theoretical model	96
18: Discussing the expected outcomes	27
19: Describing the lemma[a]	3
20: Describing the proposition	10
21: Describing the theorem	22
22: Examining the validity of the argument	1
23: Presenting the theoretical model	4
24: Referring to footnotes	10
25: Describing where to obtain detailed data	1
Total	5,763

[a] Lemma here refers to a subsidiary proposition.

DISCUSSION

Results of single-word and four-word expressions show striking differences between the researchers' extracted expressions and the automatically extracted frequency-based expressions in terms of both form and function. First, the participants included the GSL words, indicating that because the participants already knew the meaning of the words, they placed importance on the word's function, including how and where the word was used. The researchers' attention to the function of expressions is also reflected in the percentage of selected words that appear at the beginning of a sentence, many of which function adverbially. The results suggest that the researchers believe it to be helpful to master the start of the sentence to assist the flow of logic. They find one- and four-word expressions that function as linking words (adverb and adverb-based expressions and some in verb and verb-based expressions) to be useful, although not often taught in the writing classroom.

Previous studies reported non-native learners' tendencies to overuse typical transitional and linking words such as *however* and *also* as well as to underuse a variety of linking adverbial constructs (e.g., Ishikawa, 2011). Research by Ishikawa (2011) suggests that even proficient Japanese learners cannot improve their use of linking words unless they become aware of the function of such words. The selected expressions in this study imply that not all expressions are obvious linking words if taken out of context. This means that typical linking adverbials such as those examined in previous studies (Biber et al., 1999; Charles, 2011; Ishikawa, 2011) are not conclusive nor are they representative of linking expressions selected in this study. Furthermore, it is important to show a variety of linking adverbials with contextual information. As the following examples show, adverbials connect the sentences without obvious linking clues:

> . . . from a few dozen and hundreds (4, 5) to tens of thousands (6) to half a million (7). *Clearly,* the validity of such estimates is questionable.
>
> Despite these age differences in the overall level of response, children of all ages judged real and fantastical entities differently. *Theoretically,* providing a fantastical context for a novel entity should cause children to treat the entity as if it were pretend . . .

Studies such as Biber et al. (1999), Carter and McCarthy (2006), and Charles (2011) categorize linking adverbials based on function. While it is useful to understand linking adverbials, it is also possible that these categorizations would limit learners' writings, as any number of adverbs and verbs (in participle construction) can be used as linking adverbials as an author wishes, and those are the words selected by the researchers (see the following examples):

> . . . instead preceded by a time-consuming random search process. *Assuming* the Nash bargaining solution where everyone earns his or her outside options plus an equal share . . .
>
> *Following the published protocol,* (37) PCR was performed with pYT1 (A) as template.
>
> *Given the relationship between* energy investment and sustainable development and the potential direful impact that energy materials, facilities and operations can have . . .

These expressions consisted of one to several words (in this case one and four words), which can be considered to be in-between linking and non-linking words. This area seems to need some attention in teaching.

Another finding is the selection of combination words and noun phrases without post-modifier fragments. As Biber and Gray (2010) reported recently, many condensed noun phrases are used in research papers. It is difficult to extract these phrases based on frequency and range because they include author-created expressions, expressions very specific to a particular area of the discipline, and expressions using common affixes and suffixes (for vocabulary). Some combined words may already be recognized, but some appear to be entirely created by the author. The instruction in how to create noun phrases and single words by hyphenating is nonexistent in teaching. The expressions database indicates that it is important to cover this aspect of creating words and noun phrases in composition education. With this instruction, learners would realize the ability to express concepts in many ways without limiting them to certain expressions.

It could be said that the researchers who participated in this study perceive the usefulness and necessity of an expression differently from frequency-based extraction because they are both avid readers and writers—this is not the case with automated extraction. In the previous studies, many have compared language characteristics among different corpora with a different combinations of the following labels: native, nonnative (with different L1s, such as Japanese and Chinese), experts, and students (e.g., Ädel & Erman, 2012; Granger & Tyson, 1996). The results of these studies demonstrate that the native expert has greater variety and more frequent use of lexical bundles than the nonnative and native students. These studies are helpful in understanding the discrepancy between model writers and learners, providing implications for new teachings to bridge the gap. In this study, however, the nature of the selected expressions was found to be widely different from those extracted based on frequency. The difference suggests the different perspectives needed to approach mastery of language in academic writing and seems to originate with the researchers' consideration for the flow of logic, which is not manifested in the frequency of expressions, nor necessarily relative to psycholinguistic saliency (i.e., idiomaticity).

The researchers who participated in this study tended to select expressions based on certain criteria, often with the idea of a research paper in mind. It is manifested, for example, in a huge collection of "it is…." expressions. While both lists share frequent expressions, such as "it is possible that" and "it is assumed that," the researchers selected expressions that are infrequent but still useful in academic writing, such as "it is projected that" and "it is plausible to." Other expressions include hedging words in four-word expressions, as researchers find it useful to learn the function of auxiliary verbs, which are not often selected in frequency-based extractions. The four-word expressions also include "consistent with . . ." (14 expressions appear in the researcher-selected list, while none appear in the frequency-based list), which gives a description of research results, and "one" (30 expressions) as an unidentified subject in 17 instances (e.g., *one might expect that*) while "one" (13 expressions) in the frequency-based list works only as a number (e.g., *one of the major*). Similarly, only the researchers selected "let us" (12 expressions) to precede the argument. The frequency-based list, on the other hand, seems to collect similar types of expressions, represented in the concentration of noun- and preposition-based constructions as well as *the,* which comes at the beginning of the expressions.

Many of these expressions do not have meaning nor function on their own, such as in *the completion of the* and *the components of the*.

This study did not categorize the function of expressions as in previous studies. This was partly attributable to the fact that we extracted a larger number of expressions using different criteria but also to issues relating to categorization. Some researchers (Byrd & Coxhead, 2010; Liu, 2012) expressed concern that the categorization was complex, difficult, and subjective and that in many cases expressions served overlapping functions. As one participating researcher explained, expressions make sense when they are perceived in a broader context. The pedagogically motivated present research thus considers it to be more helpful to provide contextual information than meticulous functional information of expression to assist in writing.

This study is able to provide the extracted expressions with contextual information in terms of the structural unit called *move*, which, with the exception of Connor, Upton, and Kanoksilapatham (2007), was not available before. Although Cortes (2013) selected lexical bundles from a multidisciplinary introduction corpus and categorized them in terms of move, the range of expressions and contextual information was limited. The results of this study offer detailed information of the text—for example, *move 7*, which has a large number of selected expressions and is a part of the results section, together with *move 8*, which comments on the results. The concentration of expressions in certain moves means that some sections or parts of a paper are more expression-dependent than others. This information[10] cannot be provided by automated extraction, which disregards the function of the expression in relation to a whole paper but can be a great help for understanding and using the expressions in an appropriate context.

In an open-ended questionnaire, several researchers commented that, being both professional readers and writers, they understood useful expressions

[10] Concordance software such as *AntConc* (Anthony, 2011) can provide a visual map of where the target words appear within the text; however, this does not contain information on the content.

when presented in context, but it was hard for them to produce useful expressions from scratch. The expression database created in this study would not by any means replace frequency- or range-based expressions lists and their application in the classroom. However, the database is suggestive of the importance of including professional readers' and writers' diverse perspectives and changing the way expressions are presented to learners.

CONCLUSION

To investigate the kinds of expressions research specialists find useful in writing academic papers, researchers from various disciplines selected useful expressions in research papers. The researcher-extracted database revealed what frequency-based extractions were missing and the importance of researchers' multiple perspectives. For future study, it would be helpful to employ a think-aloud approach to better understand the researchers' thoughts behind why they find a given expression "useful" as well as to compare the expressions they select with those chosen by EAP practitioners. Our study's unique attempt at providing these expressions with detailed move information creates a hope to empirically examine whether they efficiently serve to construct a research paper.

The results provide important insight into what should be considered for academic writing. Useful expressions selected here include frequently used ones that overlap with those automatically extracted, expressions useful for helping the flow of logical argument, and expressions useful for explaining what an author wishes to convey. The results imply that, when teaching students to write research papers, it might be effective to show these words in context to make their forms and functions clearer, and to include surrounding sentences. The findings regarding the concentration of selected expressions in certain moves—namely, presenting results, explaining specific results, and reviewing the related literature—could lead to the efficient acquisition of those expressions, which can in turn be great tools for organizing the flow of the particular sections. The expression database may be used for developing materials, designing courses, or creating a can-do list as a reference resource. The results of our study would also answer concerns about frequency-based expressions' having low practical value because they are too obvious and easy for learners at the tertiary level. Expressions selected based on researchers' perspectives would serve to complement frequency-based expressions. The existing lists contain basic expressions for the students to learn, whereas the database created through this study could be viewed as an optional reference to further improve and enrich the students' expressions. Considering the current teaching of academic writing at research-oriented universities in Japan, the non-typical and non-obvious linking expressions as well as flexible means of creating noun expressions indicated by the researchers are particularly suggestive, which might also be applicable to other universities in an EFL context, because it would help expand students' ways of expressing themselves without limiting them to a specific set of expressions.

ACKNOWLEDGEMENTS

The data introduced in this paper were obtained with financial support from the Kyoto University President's Discretionary Funds. We are grateful to all the administrators, teachers, and students who were involved in this project. We would like to thank those researchers who participated in the study for their invaluable help and the editor and three anonymous reviewers for their useful comments on an earlier version of the paper.

THE AUTHORS

Sayako Maswana is an assistant professor at Waseda University. She is currently conducting research on genre analysis of research articles. She has published in a number of journals such as *Journal of the IATEFL ESP SIG* and *Classroom Research at the International Exchange Center, Nara Women's University*.
Email: maswana@aoni.waseda.jp

Toshiyuki Kanamaru is an assistant professor in the Department of Foreign Language Acquisition and Education at Kyoto University. He holds a Ph.D. in Human and Environmental Studies, from Kyoto University. His research interests include cognitive linguistics, natural language processing, vocabulary acquisition, and academic writing teaching.
Email: kanamaru@hi.h.kyoto-u.ac.jp

Akira Tajino, *Ph.D.*, is Professor of Educational Linguistics at Kyoto University. He is currently interested in curriculum development, EAP and pedagogical grammar. He has published articles in a number of international journals, including *ELT Journal, Journal of English for Academic Purposes*, *Language, Culture and Curriculum*, and *Language Teaching Research*.
Email: tajino.akira.5z@kyoto-u.ac.jp

REFERENCES

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes, 31*(2), 81–92.

Anthony, L. (2011). AntConc (Version 3.2.2) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263–286.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgard, & S. Oksefjell (Eds.), *Out of corpora* (pp.181–190). Amsterdam: Rodopi.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at …: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405.

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes, 9*(1)*,* 2–20.

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

Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL, 5*, 31–64.

Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English.* Cambridge: Cambridge University Press.

Charles, M. (2011). Adverbials of result: Phraseology and functions in the problem–solution pattern. *Journal of English for Academic Purposes, 10*(1), 47–60.

Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology, 14*(2), 30–49.

Chung, T. M., & Nation, I. S. P. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language, 15*(2), 103–116.

Chung, T. M., & Nation, I. S. P. (2004). Identifying technical vocabulary. *System, 32*(2)*,* 251–263.

Connor, U., Upton, T. A., & Kanoksilapatham, B. (2007). Introduction to move analysis. In D. Biber, U. Connor, & T. A. Upton (Eds.), *Discourse on the move: Using corpus analysis to describe discourse structure* (pp. 23–41). Amsterdam: John Benjamins.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*(4), 397–423.

Cortes, V. (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.

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

Farrell, P. (1990). *Vocabulary in ESP: A lexical analysis of the English of electronics and a study of semi-technical vocabulary* (CLCS Occasional Paper No. 25). Dublin: Trinity College, Centre for Language and Communication Studies.

Granger, S., & Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes, 15*(1), 17–27.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1)*,* 4–21.

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly, 41*(2), 235–253.

Ishikawa, S. (2011). A corpus-based study on Asian learners' use of English linking adverbials. *Themes in Science and Technology Education, 3*(1-2)*,* 139–157.

Kanamaru, T., Maswana, S., Sasao, Y., & Tajino, A. (2010). Muubu bunseki ni motoduku eigo ronbun hyougen deeta beesu no kaihatsu: Kyouto daigaku gakujyutu ronbun koopasu wo mochiite [Development of an English academic expression database based on move analysis: Using the Kyoto University academic paper corpus]. *Proceedings of the 16th Annual Meeting of the Association for Natural Language Processing,* 522–525.

Kanoksilapatham, B. (2005). Rhetorical structure of biochemistry research articles. *English for Specific Purposes, 24*(3), 269–292.

Liu, D. (2012). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes, 31*(1), 25–35.

Martínez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes, 28*(3), 183–198.

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*(3), 299–320.

Maswana, S. (2013). Move analysis of research papers: A collaboration with researchers from various disciplines. *Journal of the IATEFL ESP SIG, 41,*12–18.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nwogu, K. N. (1997). The medical research paper: Structure and functions. *English for Specific Purposes, 16*(2), 119–138.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*(4)*,* 487–512.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings.* Cambridge: Cambridge University Press.

Swales, J. M. (2004). *Research genres: Explorations and applications.* Cambridge: Cambridge University Press.

Tajino, A., & Kanamaru, T. (2011). An interdisciplinary data-based academic word list: Developing an EAP curriculum. *IATEFL 2010 Harrogate Conference Selections*, 96–97.

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

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