Lexical Bundles in Learner Writing: An Analysis of Formulaic Language in the ALESS Learner Corpus

David ALLEN

Abstract

Lexical bundles are empirically derived formulaic units of language which are register-specific and perform a variety of discourse functions. Because these units of language contribute to the linguistic make up of specific registers, they can be important indicators for determining the success of language users within these discourse communities. Therefore, language learners need to assimilate appropriate use of lexical bundles in order to create effective and successful, register convergent discourse. The present research examines the type and frequency of lexical bundles in the ALESS Learner Corpus. Three analyses are performed focusing on accuracy, grammatical class and function of lexical bundles in learner writing, and the findings are compared with other corpora in order to assess convergence of learner use with native speaker use and published academic writing. The findings reveal a number of areas where pedagogical applications and further research are recommended.

Introduction

1.1. Lexical Bundles

Lexical bundles, defined as 'the most frequently occurring lexical sequences in a register' (Biber, Conrad & Cortes, 2004; Biber, Johansson, Leech, Conrad, & Finegan, 1999) have received increasing attention over the last ten years. Lexical bundles, also referred to as N-grams, often do not fit with traditionally idealized units of language, but may cross over a number of structures e.g. *In this study we, should be noted that*. Nevertheless, because lexical bundles are discovered purely on the basis of their frequency within texts, they may be considered empirically derived units. Furthermore, these bundles often have discernible functions within a particular register. For example, Biber, Conrad and Cortes (2004) found that classroom teaching tends to feature more personal stance bundles (e.g. *you have to do*) than academic prose, which in contrast features more impersonal stance bundles (e.g. *it is necessary to*). Also, discourse organising bundles were found to differ across spoken and written registers: *if you look at* was a common bundle found in the former, while *on the other hand*, is found more often in the latter. In his monograph entitled *University Language*, Douglas Biber (2006) notes that 'lexical bundles are crucially important for the construction of discourse in all university registers' (p.174). Therefore, for learners of English for Academic Purposes, or indeed for learners within any of the more specialised academic discourse communities, knowledge and use of lexical bundles must be equally important.

For language learners, the knowledge and use of a wide range of formulaic language helps them to achieve naturalness in language use. Conversely, misuse of formulaic language has been shown to be a potential source of communication difficulties (Millar, 2009). Using eye-tracking methodology, Millar measured native speaker reading times of collocations that had been taken from Japanese learners' academic writing and their native speaker equivalents. The findings indicate that learner collocations, that is, those which are divergent from native speaker norms, take longer to process when reading.

Michael Hoey (2005) discusses this issue in terms of lexical primings. Continual exposure to language reinforces the associations which words have, including semantic, collocational, colligational and prosodic associations. Though all language users' primings will differ (due to the variation of language which one is exposed to), there emerge patterns of language use and levels of acceptability within specific registers. These primings thus determine and propagate formulaic language such as lexical bundles. Ken Hyland makes the following observation about lexical bundles:

> (they) are familiar to writers and readers who regularly participate in a particular discourse, their very 'naturalness' signalling competent participation in a given community. Conversely, the absence of such clusters might reveal the lack of fluency of a novice or newcomer to that community (2008, p.2).

Yet, learners rarely have competent use of such lexical bundles when they begin to study academic discourse in a second language, even if they have experience of participation in such communities in their first language. Research has shown that learners of English from a particular language group produce language features in their writing which differs from native speaker norms (Altenberg & Granger, 2001; Hyland & Milton, 1997). These analyses, termed *contrastive interlanguage analysis* (Granger, 2002), have shown that in many cases, language learners overuse, underuse or misuse particular language functions and exponents. For example, Chinese L1 learners of English differ in their use of hedging and boosting in academic writing (Hyland & Milton, 1997). They tend to underuse hedging in their academic essays leading to a more authoritative tone than found in native speaker essays (*ibid.*).

Lexical bundles have also been investigated in learner writing (Cortes, 2002; Rica-Peromingo, 2009). By comparing the findings of an undergraduate Spanish learner corpus with a corpus of American university students and another of professional native speaker writers, Rica-Peromingo (2009) demonstrated that there were significant differences in the type and frequency of learners' use of lexical bundles. Learners tended to over- and underuse particular language units, such as linking and stance adverbials, in ways divergent from native speaker writers.

The aims of the current paper are to investigate learners' use of formulaic language in their written production of science research articles. These findings will be compared to results from reference corpora and other recent studies (e.g. Biber, Conrad & Cortes, 2004; Hyland, 2008), and the degree of convergence between learners' production and published and/or native speaker writing will be evaluated. Where learners' formulaic language diverges from that typically found in the target genre, suggestions will be made as to the reasons for the divergence, and also for teaching applications. In the following section I will outline the use of corpora in the analysis of formulaic language. I will then describe the current corpus and also the reference corpora used in the study, before presenting the findings of the analysis.

2. Methods and Materials

2.1. Corpora

Much of the work discussed in the previous sections is based upon research using corpora. The use of large-scale computerised corpora has made the study of formulaic language possible, without which only relatively small texts or parts of texts could be analyzed manually. One form of specialised corpora which has received considerable attention is the learner corpus (Gilquin, Granger, & Paquot, 2007; Granger, 2002). By compiling written documents such as student essays and reports, teachers and researchers can create a rich dataset with which various analyses can be conducted. Examples of well-known learner projects are the International Corpus of Learner English (ICLE), the Japanese English as a Foreign Language Learner (JEFLL) corpus and the Cambridge Learner Corpus (CLC). In the following section I will describe the learner corpus constructed to represent learner production on a scientific writing course at the University of Tokyo.

2.2. The ALESS Corpus

The Active Learning of English for Science Students (ALESS) course runs for all first-year undergraduate science majors at the University of Tokyo. The current ALESS learner corpus consists of 847 final research papers collected over one semester[1]. These papers are the end-product of a writing programme which emphasises the process of writing, particularly in the form of peer-review and revision of texts. The reports are written in the IMRD format, typical of the research article genre, with a focus on the rhetorical structure of these articles; each paper includes an abstract, introduction, method, results, discussion (and conclusion) and references section. Some also contain appendices. The target register is broadly written academic English in the sciences, and is relatively broad in scope as students are yet to specialise in their chosen subjects. All students were asked for permission to use their work for the creation of the current corpus; of a total of 990 submissions 143 (14%) declined leaving the total number of 847 individual files. Each file has information including student name, class, instructor, and date of submission. This information is confidential and is omitted from public documentation. The number of total words in the corpus is 731, 612. The concordancer Antconc was used for the present investigation (Anthony, 2006).

2.3 Reference Corpora

In the following sections the analyses will use a number of other corpora for reference. The purpose of reference corpora is to provide comparison of the rates of occurrences of certain language features, in this case lexical bundles, with the language use of other populations, in this case those of published research article authors and native speaker writers in similar contexts. I describe here three corpora which will be referenced at various stages in the current analysis. The Professional English Research Consortium (PERC) Corpus is currently available online (see http://www.corpora.jp/~perc04/) as a 17-million word corpus of published research articles divided into 22 domains of academic disciplines, such as Computer Science, Biology and Medicine. Another useful reference corpus is the British Academic Written English (BAWE) corpus. The BAWE contains 3000 texts of student writing, from a variety of genres and text types, totaling 6.5 million words. Finally, Ken Hyland's (2008) corpus of 3.5 million words of academic written English includes research articles, masters' dissertations and doctoral theses[2]. The articles are all published in leading journals while the theses and dissertations were written by Cantonese L1 writers studying in Hong Kong universities. Four disciplines are represented: Electrical Engineering, Biology, Business Studies and Applied Linguistics.

3. Findings and Discussion

Using the concordancer, all 4-word lexical bundles were computed. Those occurring 40 times or more were kept for analysis giving a total of 144 lexical bundles (see Appendix 1 for full list including rank and frequency information). The decisions to examine only 4-word bundles and to set the frequency cut-off point at 40 or less are both motivated by the need to limit the amount of data gained for the current analysis.

The results are divided into three main sections: Firstly, I will consider the accuracy of the bundles (as this is a learner corpus not all bundles will necessarily be grammatically accurate); secondly, I will classify the bundles by grammatical structure and describe how well these findings converge with the target register norms; finally, I will present a functional analysis of the items and comment on the convergence in terms of the use of bundles in student writing. Frequency figures are given in brackets following the examples of lexical bundles; rate of occurrence per million words is also given for comparison purposes (/1M).

3.1. Accuracy

An initial observation is that there are apparently no high frequency bundles which contain grammatical errors; in other words, grammatical accuracy is high. This is most likely a product of the considerable revision and editing that student writers undertake in the form of peer review, peer conferencing and individual review. An exception, however, may be the use of the bundles surrounding the stem *result of*, as highlighted by *result of this experiment* (65; 89/1M) and *result of the experiment* (45; 62/1M). These bundles are frequent in the ALESS corpus but not so in the PERC corpus (1; 0.06/1M) or the BAWE (2; 0.31/1M). Closer inspection of these bundles by analyzing the key word in context (KWIC) concordances reveals regular misuse, as illustrated in the examples below:

> *The* ***result of this experiment*** *was expressed by following graphs.*
>
> *The weakness of the experiment is difference among individual's ability to learn something by heart and that subjects are only 30, so the* ***result of this experiment*** *may include errors.*

The singular form of *result* creates a subtle conflict in these instances. Scientific investigation by way of experiments, calculations and analyses tends to produce more than one result; it is rare to have results from experiments reported in a single figure, or in the form of a simple yes/no answer. The reason why this conflict in usage is indeed subtle is because of other uses of the stem *result of*, as demonstrated below:

1. (X) (*to be*) ***a/the result of*** (Y)
2. As ***a result of*** (X), (Y)

These forms, which utilise the singular form of *result*, are highly frequent bundles found in scientific discourse (Hyland, 2008; Biber et al, 1999). The first use co-occurs with the verb *to be* and controls two noun phrases. The adverbial conjunctive *as a result of* is particularly common in the natural sciences, but is also frequently used in the softer sciences (Biber, 2006, p. 167). Although these forms are found in the ALESS corpus, they are less frequent than the misused bundles. In fact, there is considerable evidence of misuse of these forms also, as indicated by the confused example below:

> ***As a result of this experiment****, the average time of each color was as follows; black cloth took 76minutes, blue 82minutes, green 86minutes, red 93minutes, yellow 101minutes and white 104minutes.*

The high occurrence of bundles surrounding the stem *result of* in academic literature may account in part for the learners' use of the singular form, if the differences in usage are not noticed in the input. However, a more convincing explanation for the misuse may be found in first language transfer. The number system of English is notably difficult for Japanese learners; the surface form 結果 (kekka – result) in Japanese can denote both single and multiple findings, whereas in English this difference is expressed by the use of the singular/plural forms. I suspect a combination of transfer and input factors to be the source of learners' errors in this case. Nonetheless, given the inconsistent levels of accuracy concerning the use these lexical bundles, it may be prudent to devise exercises to highlight the various uses of these forms. This could be achieved using example concordance lines as provided by the BAWE and ALESS corpora, with which learners extend their knowledge of both accurate and inaccurate uses of the forms.

3.2. Grammatical structure of bundles

Lexical bundles were classified into grammatical categories based on similar analyses by Hyland (2008) and Biber et al (1999). The full listings are presented in Appendix 2 and a breakdown of the categories is presented below in Table 1; the main findings are discussed below.

Structure	**Number**	%
NP + of	36	41.37
Other NPs	4	4.60
Prepositional Phrase + of	3	3.45
Other Prepositional Phrase	14	16.09
Passive + Prepositional *Phrase/ That*-Complement	5	5.75
Anticipatory *it* + V/Adj	7	8.05
Be + N/Adj Phrase	2	3.30
Others	16	18.39
Totals	**87**	**100**

Table 1: Grammatical Categories of Lexical Bundles

3.2.1 Noun Phrases

The largest grammatical category of lexical bundles is the noun phrase (NP) + *of* structure, making up 41% of the total number of bundles in the analysis e.g. *the temperature of the, the length of the, the purpose of the*. Given that the register is written academic English, specifically scientific research papers, this result is not particularly surprising. The register of the science research article seeks to make claims of factuality based on a full, explicit methodology and transparent analysis. In the aim of replicability, science articles contain considerable detail, which is often compressed in the complex noun phrase (Biber et al, 1999). In general, this finding demonstrates that the learners are producing written prose characteristic of the target register, at least in terms of this grammatical category. However, the percentage occurrence of these NP bundles (41%) is notably high even when compared to the Biology (23.7%) and Electrical Engineering (22.3%) sub-corpora used in Hyland's (2008) study. Whether or not this reveals overuse by learners is a question which requires a more extensive analysis of noun phrases within the corpus. Anecdotal evidence would suggest that constructing noun phrases is a problematic area for learners, particularly in deciding whether to use *NP + of* structures, *N + N* sequences or the genitive (*'s*). It may be that learners require additional support in this area of academic writing.

3.2.2 Passive forms

Another finding worthy of discussion is the percentage of passives + prepositional phrases. While this grammatical class of bundle accounts for only 6% of lexical bundles in the ALESS corpus, this figure rises to around 30% in Hyland's corpus (2008). Although these forms may certainly be found in the ALESS corpus, e.g. *can be said that* and particularly in the anticipatory-*it* structures e.g. *it is well known, is known that the*, they appear underrepresented. Passives are common in academic written discourse for a number of reasons documented elsewhere (e.g. Biber et al, 1999). In terms of lexical bundles, the examples cited above may serve particular functions in the texts, and it may be these functions which are underrepresented; this issue is discussed in the following sections where the focus is the functions of lexical bundles.

3.3. Functions of Lexical Bundles

It has been shown in previous studies (Biber, Conrad & Cortes, 2004; Biber et al., 1999; Hyland, 2008) that lexical bundles tend to have functional characteristics that are representative of the register in which they are found. Following the classification systems used in these previous studies I have categorised the most common lexical bundles in the ALESS corpus by their functional role (see Appendix 3 or the full list). The three main categories for functional items identified by Hyland (2008, p. 13–4) have been used here:

- Research-oriented bundles – 'help writers to structure their activities and experiences of the real word'
- Text-oriented bundles – 'concerned with the organisation of the text and its meaning as a message or argument'
- Participant-oriented bundles – are 'focused on the reader or writer of the text' (*ibid*, p.14)

Each of these main functional types can be further subdivided into more specific functional roles as is shown in table 2.

Type of Bundle	Number	% of Total
Research-Oriented Bundles	**58**	**66.67**
Location	5	5.74
Procedure	11	12.64
Quantification	9	10.35
Description	22	25.29
Topic	4	4.60
Relational Bundles	7	8.05
Text-Oriented Bundles	**19**	**21.84**
Transition Signals	2	2.3
Resultative Signals	15	17.24
Structuring signals	0	0
Framing Signals	2	2.30
Participant-Oriented Bundles	**10**	**10.35**
Stance features	7	8.05
Engagement features	2	2.30
Non-classifiable	**1**	**1.14**
Total	**87**	**100**

Table 2: Lexical Bundles by Functional Type

3.3.1 Research Oriented Lexical Bundles

Research-oriented bundles tend to dominate scientific discourse in the ALESS corpus, similarly to the findings from other science and technology corpora (Hyland, 2008). This is because of the need to relay detailed information about the research, so much so as to make any methodology replicable and in order to persuade the reader of the precision and validity of the findings recorded. In the hard sciences the importance of relaying empirical evidence is central to any research article; Hyland (2008, p.15) succinctly summarises this epistemological framework as:

> ‘an ideology which emphasises the empirical over the interpretive, minimizing the presence of the researchers and contributing to the “strong” claims of sciences’ (original emphasis).

These research-oriented bundles found in the ALESS corpus describe:

- location – indicating time and place e.g. *in this study I, in this experiment the*
- procedure – indicating methodology or purpose of the research e.g. *the purpose of this, the experiment was conducted*
- quantification – describing amount or number e.g. *the amount of water, is one of the, the number of the*
- description – detailing qualities or properties of materials e.g. *the temperature of the, the length of the, the surface of the*
- topic – by being subject-specific and focused e.g. *available at http www, the growth of plants, http ja Wikipedia org*
- relations – indicating relationships or contrasts between materials or number e.g. *the relation between the, the proportion to the, the difference of the*

The examples given in italics above are the most common bundles in the ALESS corpus from each classification. A number of findings require expanding. Firstly, for the procedure bundles, there is a high occurrence of strings which overlap around the common bundle *the purpose of this research is to*. These include: *The purpose of this* (129, 176/1M), *purpose of this research* (86, 118/1M), *of this research is* (75, 103/1M), *this research is to* (53, 72/1M) and *the purpose of the* (55, 75/1M). It has been argued, however, that this bundle is more common in written academic journals from disciplines such as applied linguistics and business studies, and is rarely found in the research of the hard sciences (Myers, 1992:304). Similarly, Hyland (2008) found the bundles *the purpose of this/purpose of this research* to occur frequently in his applied linguistics sub-corpus, but not in the hard sciences or engineering sub-corpora. However, the bundle *purpose of this research* does occur seven times in the PERC corpus (0.43/1M) within the disciplines of agriculture, medicine, physics and computer science. This last finding appears to limit the bundle to low frequency in the sciences, but not to inappropriacy. The practical value may thus be advisable to provide a greater range of expressions for learners so to aid convergence upon these norms; alternatively, it may be better to show how the purpose of research is often expressed together with the main findings of the research, using such phrases as *here we show* (see Allen and Middleton, *submit-*

ted).

A second point of interest is the abundance of bundles of description which often follow the *NP + of* structure, as shown below:

the **strength** of the
the **height** of the
the **average** of the
the **shape** of the
the **density** of the
the **volume** of the
the **mass** of the
the **concentration** of the

As was noted in section 3.2.1, the *NP + of* construction is a common characteristic of the academic, and particularly the scientific written register. Of these constructions, those which describe mathematical relationships (*the average of the*), the dimensions (*the height of the*) or qualities (*the concentration of the*) of materials form the single largest classification in the ALESS corpus (25.29%). This finding indicates that the apprentice writers in this study focus considerable effort describing their experiment. Certainly, this feature of writing is found primarily in the method sections of the reports, but such bundles can also form part of the topic of the experiment, for example, *the strength of materials* can form the basic theme of many simple experiments.

The central themes in the corpus are highlighted by topic bundles, and are typically corpus-specific. One of the recurring themes in the ALESS corpus is *the growth of plants*, which is possibly the most often chosen topic of investigation. Unfortunately, the common occurrence of *http ja Wikipedia org* highlights other difficulties in research skills and library skills faced by language learners. Developing such skills may need to become a more integrated part of writing pedagogy and may certainly be the key to improving background research sections in written reports (Yen, 2008).

3.3.2 Text-oriented bundles

Many of the text-oriented bundles are cohesive and aim to develop arguments by making logical connections between propositions; these are classified into the following types:

- Transition signals – signal cohesive relations in discourse, such as contrastive phrases: *on the other hand, the other hand the*

- Framing signals – serve to frame an argument by limiting its conditions: *in the case of, in the same way*
- Structuring signals – are used to structure larger sections of discourse and may include text deictics (no occurrences in the corpus)
- Resultative signals – signal results and consequences of actions or events: *the result of this, the effect of the, I found that the*

The most frequent bundle in the whole corpus is the transition signal, *on the other hand*, a finding which converges with the academic prose register norms (Biber, 2006; Hyland, 2008). This bundle is very typical of written academic discourse, but is less frequent in spoken genres. Furthermore, the two framing signals, *in the case of* and *in the same way*, identified in the ALESS corpus are also highly frequent in academic written texts (e.g. Hyland, 2008). Yet, the lack of any structuring signals, such as *in the next section* and *as can be seen*, is surprising. A possible explanation for this is the limited length of the reports which learners produce; for shorter research papers it is not common to find such structuring devices. For example, in Nature's former Brief Communications papers, which form part of a 'shorter communications' genre (Swales, 2004), such signals are rarely seen. Another function of structuring signals is to guide the reader to figures and tables presented in the paper, and these could be expected to be more common as almost all ALESS papers include such visual descriptors. As noted in section 3.2.2, the underrepresentation of passive forms appears congruent with this lack of highly frequent structuring bundles in the texts. Further research focusing on the range and frequency of passives, and their functions in the papers, may shed light on this issue.

Whereas some text-oriented signals appear to be poorly represented, resultative signals are conversely very frequent. Bundles such as *these results indicate that* and *the effect of the* are highly frequent in the ALESS corpus, and are also typical of science writing. As the aim of the hard sciences is primarily to present results, thereby making a claim, resultative signals are naturally prominent in the discourse. The range of these bundles in the corpus show that the structure *the result(s) of this experiment V + that* is highly frequent, as was discussed in regard to grammatical structure of bundles in section 3.1.

3.3.3 Participant-oriented bundles

The final class, participant-oriented bundles, relates to either the reader or the writer and include:

- Stance features – indicating the writers position (e.g. hedging bundles): *can be said that, it is widely known, it is known that*
- Engagement features – indicating the writers attempts to engage the reader in the discourse process: *it is difficult to, it is necessary to*

In the ALESS corpus the most highly frequent stance bundles are those expressing epistemic stance e.g. *it is known that* (225, 308pmw), *is widely known that* (56/77pmw) and almost all include the verb *to know* in its past participle form, *known*. These bundles express the writer's position in that he or she is presenting information as accepted fact. If, for instance, the writer did not feel information was accepted, other forms utilising reporting verbs could be used, e.g. *it has been suggested, some have argued*. The highly frequent use of the verb *known* in these bundles is suggestive of an overreliance on the structure *It is* (*widely/well*) *known that*. In fact, in the papers this structure forms one half of a two-part rhetorical device typically found in research article introductions. In the introductory sections of research papers, the writer seeks to create a research space (CARS; Swales, 2004) by presenting background information of an area, specifically leading the reader to a gap in the research. The bundle above forms a simplistic opening structure for the 'known' background before identifying the 'unknown' i.e. the gap. It is very likely that learners require a greater range of expressions to achieve these aims.

Other stance bundles *it can be said* and *can be said that*, are both hedging structures which modify (weaken) a claim. It has been noted that stance in academic writing is often expressed by hedging (Hyland, 1996; 1994). However, linguistic transfer may also help to explain the high frequency of the these bundles. As the L1 is Japanese, *it can said that* is often expressed by the form と言われる (*to iwareru*). Although in Japanese the form is somewhat ambiguous as the suffix-れる (*-reru*) can be used to express the passive *was said* or the possible *can say*, it is more likely the repeated occurrence of this form within the academic written registers in Japanese which is responsible for its transfer here. Lexical primings in one language may well be facilitative in writing in a second language, and if structures co-occur in two languages and in the same register, these are strong candidates for transfer; a similar argument was put forward by Rica-Peromingo (2009) concerning the transfer of lexical bundles from L1 Spanish to L2 English.

Engagement features achieve the aim of engaging the reader. Note that in less formal texts this would be achieved by the use of phrases such as, *and you have to, we're going to do* (Biber, Conrad & Cortes, 2004). In impersonal writing though, other forms are preferred, such as *it is necessary to, it is difficult to*. Though there are only two highly frequent engagement bundles both of these are highly frequent in other academic corpora (e.g. Biber, Conrad & Cortes, 2004; Biber et al., 1999).

4. Conclusions

Based on the findings presented here there appears to be considerable convergence between the lexical bundles employed in student writing in the ALESS corpus, when compared with published and native speaker writing. Undoubtedly the high level of accuracy and appropriacy is partly due to the continual process of revising and editing texts which the writers perform. Learners' successful adoption of register-convergent bundles should be encouraged through highlighting appropriate use in texts and similarly noticing divergent use, such as those used in spoken academic discourse.

In the analysis a number of areas for of pedagogical application were noted. Firstly, certain bundles such as those including the stem *result of* may be suitable items with which to design learning activities using concordances as shown in section 3.1. Also, further research into the use of noun phrase constructions, will reveal whether learners are overusing *NP + of* constructions over other alternatives and whether this is due to difficulties in constructing noun phrases appropriately in science discourse. The acquisition and use of noun phrases has previously been shown to be problematic for Japanese learners of English, not least because of the different patterns of modification across the two languages (Miura, 2008). Such findings would merit the production of suitable learning activities to promote use of a range of noun phrase structures in learner writing. Finally, learners' use of passive structures, such as structuring signals, may warrant further investigation to determine why there is a possible underuse of these forms.

The use of learner corpora as a basis for designing learner materials is widely advocated in the field (Gilquin, Granger & Paquot, 2007: 332) and the ALESS learner corpus has been shown to be a valuable source of data for researching learner language production. Flowerdew (2001) noted that learner corpora studies have tended to have less pedagogical impact than native speaker corpus studies. However, by comparing learner corpora with reference corpora, language production can be investigated and when such production diverges from the target register, teaching materials can be produced specifically targeting the learner population, in this case undergraduate Japanese learners of English studying science subjects.

A final note should address the range of lexical bundles analysed in this research. It is clear from even a cursory glance that lexical bundles overlap in form e.g. *the purpose of this, purpose of this research, of this research is to*. However, this is due to the empirical methodology employed in extracting all n-grams of predefined length from text corpora and is not in the least detrimental to the value of the findings. Formulaic language is not always easy to categorise into neat two, three or four word chunks, in fact linguists have struggled for decades to draw boundaries upon the varying degrees of formulaicity in language. The beauty of lexical bundles is that they are decided based on frequency alone, bypassing much of this messy quantification. However, there remains a level of subjectivity in deciding which lexical bundles are suitable to highlight in the classroom and which are less so. It is thus the educator's duty to devise effective ways of teaching the most important of these highly productive, register-specific bundles of words.

Notes

1. The corpus is an ongoing project and is expected to expand each year

with a roughly similar number of texts added each semester as make up the present corpus.

2. Ken Hyland's corpus is not freely available for use; however, the findings of Hyland (2008), which is based on research using this corpus, will be referred to in the present study.

References

Allen, D.B. and Middleton, G., (2009). Abstroductions: The Compression of Part-Genres in *Nature*'s Brief Communications. *Submitted*.

Anthony, L. (2006). AntConc 3.1.302 (Windows). Waseda University: Freeware.

Altenberg B. and Granger S. (2001). The grammatical and lexical patterning of *make* in native and non-native student writing. *Applied Linguistics 22(2), 173–194*.

Biber, D. (2006). *University Language: A corpus-based study of spoken and written registers.* Amsterdam: John Benjamins.

Biber, D., Conrad, S. and Cortes, V. (2004). 'If you look at . . . .: Lexical Bundles in University Teaching and Textbooks'. *Applied Linguistics 25 (3), 371–405.*

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999). *Longman Grammar of Spoken and Written English.* Longman, Harlow.

Cortes, V. (2002). Lexical bundles in Freshman composition. In R. Reppen, S. M. Fitzmaurice, & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 131–145). Amsterdam: Benjamins.

Flowerdew, L. (2001). The exploitation of small learner corpora in EAP materials design. In M. Ghadessy, & R. Roseberry (Eds.), *Small corpus studies and ELT* (pp. 363–379). Amsterdam: Benjamins.

Gilquin, G., Granger, S. and Paquot, M. (2007). Learner Corpora: The Missing Link in EAP Pedagogy. *English for Academic Purposes, 6, 319–355.*

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung and S. Petch-Tyson (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* (pp.38–51) Lund: Lund University.

Hoey, M. (2006) *Lexical Primings: A new theory of words and language.* Oxon: Routledge.

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

Hyland, K. (1996). Writing without conviction? Hedging in science research articles. *Applied Linguistics*, *17, 433–454*.

Hyland, K. (1994) Hedging in Academic Writing and EAP Textbooks. *English for Specific Purposes, 13 (3) 239–256.*

Millar, N. (2009) *Assessing the processing demands of learner collocation errors.* Poster presented at Corpus Linguistics 2009, Liverpool, U.K.

Myers, G. (1992). 'In this paper we report . . . ': Speech acts and scientific facts. *Journal of Pragmatics, 17, 225–313.*

Rica-Peromingo, J.P. (2009) The use of lexical bundles in the written production of Spanish EFL university students. *Applied Linguistics for Specialised Discourse. Conference Proceedings.* Riga: University of Latvia Publishing. (pp 1–7).

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

Yen, S.C. (2008). Teaching EFL Undergraduates Research Writing through Modeling: Students' Difficulties, Perception and Their Use of Rhetorical Patterns in the Introduction Section. *25th International Conference of English Teaching and Learning 2008 Conference on English Instruction and Assessment.*

Miura, A.（三浦愛香）(2008). 会話 (NICT JLE) vs. 作文 (JEFLL) コーパスの比較と分析：英語学習階段と名詞の内部構造発達．[A comparison and analysis of spoken and written corpora: English learner proficiency and noun phrase construction] 英語コーパス研究第15号抜刷英語コーパス学会．[English Corpus Research, Issue 15. Japanese Association for English Corpus Studies].

Acknowledgements

I wish to thank to Moe Sakai for her immense contribution in creating and maintaining the student database, without which the construction of the corpus would have pained me much more.

Appendix 1

All lexical bundles occurring 40 times or more

Rank	**Frequency**	
1	288	on the other hand
2	246	the temperature of the
3	235	the relation between the
4	225	it is known that
5	222	the relationship between the
6	191	the amount of water
7	181	in this study i
8	178	in proportion to the
9	154	in this experiment i
10	129	the purpose of this
11	121	the length of the
12	118	i found that the
13	118	in this experiment the
14	113	the difference of the
15	103	the result of this
16	101	and the amount of
17	101	in the case of
18	97	is one of the
19	97	the number of the
20	91	the amount of the
21	90	the temperature of water
22	89	available at http www
23	88	the surface of the
24	86	purpose of this research
25	86	the effect of the
26	85	the weight of the
27	85	used in this experiment

28	84	the center of gravity
29	84	the experiment showed that
30	81	can be applied to
31	81	the height of the
32	81	the strength of the
33	80	the change of the
34	80	the result of the
35	79	the results of this
36	78	the other hand the
37	77	at the same time
38	77	the growth of plants
39	76	it can be said
40	76	temperature of the water
41	76	the results showed that
42	75	of this research is
43	74	can be said that
44	74	the color of the
45	73	it is difficult to
46	73	it is necessary to
47	70	the center of the
48	70	the results of the
49	70	these results indicate that
50	69	the average of the
51	68	that the amount of
52	67	the shape of the
53	66	is in proportion to
54	65	result of this experiment
55	64	of the amount of
56	64	the density of the
57	63	the size of the
58	60	and the number of
59	59	http ja wikipedia org
60	59	ja wikipedia org wiki
61	59	the difference between the
62	59	the volume of the
63	58	in order to test
64	58	of this experiment is
65	58	the experiment was conducted
66	56	in this research i
67	56	is proportional to the
68	56	is widely known that
69	56	it is widely known
70	56	results of this experiment
71	55	the mass of the
72	55	the purpose of the
73	54	this study i found
74	53	this research is to
75	52	in order to make
76	52	in the same way
77	52	it is well known
78	52	to the amount of
79	50	in this research the
80	50	is known that the
81	50	it was found that
82	50	on the surface of
83	50	study i found that
84	50	that there was a
85	50	the concentration of the
86	50	the speed of the
87	50	the surface area of
88	49	experiment showed that the
89	49	the influence of the
90	48	is well known that
91	48	of white radish sprouts
92	48	than that of the
93	48	that there is a
94	48	the reason for this
95	48	the surface tension of
96	47	on the growth of
97	47	the fact that the
98	46	however it is not
99	46	my findings indicate that
100	46	the angle of the
101	46	the results show that
102	46	this experiment was conducted
103	45	in this paper i
104	45	it is not known
105	45	it is possible to
106	45	result of the experi-

		ment
107	45	the diameter of the
108	44	higher than that of
109	44	the amount of salt
110	43	between the amount of
111	43	i measured the time
112	43	in inverse proportion to
113	43	the freezing point of
114	43	the reason why the
115	43	the results suggest that
116	43	the same amount of
117	43	this result indicates that
118	43	used in this research
119	42	be said that the
120	42	in order to investigate
121	42	it is said that
122	42	to test whether the
123	41	g g g g
124	41	i j image i
125	41	image i j image
126	41	j image i j
127	41	lower than that of
128	41	of the present research
129	41	of water in the
130	41	the coefficient of restitution
131	40	after and to minutes
132	40	after to minutes after
133	40	and to minutes after
134	40	as the amount of
135	40	for a long time
136	40	minutes after and to
137	40	minutes after to minutes
138	40	minutes after when the
139	40	of the center of
140	40	purpose of the present
141	40	the distance between the
142	40	to minutes after and
143	40	to minutes after when
144	40	when the amount of

Appendix 2

Lexical bundles classified by grammatical category

NP + of

the amount of water
the number of the
the amount of the
the purpose of this
purpose of this research
the purpose of the
the temperature of the
the length of the
the result of the
the temperature of water
the surface of the
the weight of the
the center of gravity
the height of the
the strength of the
the change of the
temperature of the water
the color of the
the center of the
the average of the
the shape of the
the density of the
the size of the
the volume of the
the mass of the
the concentration of the
the speed of the
the surface area of
the growth of plants
the difference of the
the effect of the
result of this
the results of this
the results of the

result of this experiment
results of this experiment

Other NPs

the relation between the
the relationship between the
the difference between the
the other hand the

PP + of

to the amount of
on the surface of
in the case of

Other PP

on the other hand
in this study i
in this experiment i
in this experiment the
at the same time
in this research i
in proportion to the
in the same way
of the amount of
in order to make
in this research the
of this research is
in order to test
of this experiment is

Passive + PP / *That*-Comp

can be said that
Passive VP + That
is widely known that
is known that the
used in this experiment
can be applied to

Anticipatory *it* + V/Adj

it is known that
it can be said
it is widely known
it is well known
it was found that
it is difficult to
it is necessary to

Be + N/Adj Phrase

is one of the
is proportional to the

Others

i found that the
the experiment showedthat
the results showed that
these results indicate that
the experiment was conducted
is in proportion to
study i found that
this study i found
this research is to
that there was a
that the amount of

Appendix 3

Lexical Bundles classified by function

Research-Oriented Location

in this study i
in this experiment i
in this experiment the
at the same time
in this research i

Procedure

the purpose of this
purpose of this research
used in this experiment
of this research is
in order to test
of this experiment is
the experiment was conducted
the purpose of the
this research is to
in order to make
in this research the

Quantification

the amount of water
and the amount of
is one of the
the number of the
the amount of the
that the amount of
of the amount of
and the number of
to the amount of

Description

the temperature of the
the length of the
the temperature of water
the surface of the
the weight of the
the center of gravity
the height of the
the strength of the
the change of the
temperature of the water
the color of the
the center of the
the average of the
the shape of the
the density of the
the size of the
the volume of the
the mass of the
on the surface of
the concentration of the
the speed of the
the surface area of

Topic

available at http www
the growth of plants
http ja wikipedia org
ja wikipedia org wiki

Relational

the relation between the
the relationship between the
in proportion to the
the difference of the
is in proportion to
the difference between the
is proportional to the

Text-Oriented Transition Signals

on the other hand
the other hand the

Resultative Signals

i found that the
the result of this
the effect of the
the experiment showed that
can be applied to
the result of the
the results of this
the results showed that
the results of the
these results indicate that
result of this experiment
results of this experiment
this study i found
it was found that
study i found that

Structuring Signals

None

Framing Signals

in the case of
in the same way

Participant-Oriented Stance Features

it is known that
it can be said
can be said that
is widely known that
it is widely known
it is well known
is known that the

Engagement Features

it is difficult to
it is necessary to

Non-classifiable

that there was a