Data Driven Learning by Discovering Lexical Bundles Using Corpus Resources

Kamal Yusuf
Faculty of Adab and Humanities
UIN SunanAmpel Surabaya
e-mail: kamalyusuf@uinsby.ac.id

***Abstract*—Lexical bundles have recently commenced to attract huge attention in corpus-based research. Knowledge of lexical bundles is essential for students learning foreign language, especially those with English for specific purposes. By employing data driven learning, students are able to lead their awareness and learning autonomy in guiding their language discovery tasks. This paper showcases the facets and usage of Corpus of Contemporary American English (COCA) and Lexical Tutor for English vocabulary teaching and learning. In addition, this article provides introductory description of a practical corpus tool for discovering English lexical bundles and how integrating the corpus in teaching for boosting students' vocabulary. The main features of the tool such as N-gram underpins in this paper.**

Keywords: lexical bundles, vocabulary, corpus linguistics

I. INTRODUCTION

Learning foreign languages such as English is an activity that is quite challenging for many Indonesian students. In addition to learning its grammar that is different from their mother tongue, one that is quite thorny is the vocabulary. Vocabulary is one of the most important aspects. Several studies show that vocabulary is one of the basics in learning foreign languages as it is central of a language. Students are directed to learn word by word from a language vocabulary.

Furthermore, researchers have begun to broaden their attention in a group of words. A group of words is a sequence of two or more words that frequently occur in text. This sequence of words by several researchers is called with different names, such as, logical sentence stem (Pawley and Syder, 1983), clusters (Scott, 1997), or N-grams (Stubbs, 2005), phraseology (Granger &Meunier, 2008), formulaic language (Schmitt & Carter, 2004; Wray, 2008) and lexical bundles (Biber& Conrad, 1999). All the same, they principally refer to the same notion.

Some studies demonstrated the importance of knowing lexical bundles (LB). The studies proved that knowledge of LB reveals a higher level of language abilities than knowledge of individual words. Meanwhile, learners with lower language ability rely more on individual words (Vidacovic and Barker, 2010). This fact shows how important the knowledge of LB is in language learning as they can play a focal point in building language fluency and confidence.

II. LEXICAL BUNDLES ACROSS ACADEMIC

Jespersen (1924) and Firth (1957) are considered as the early scholars who introduced the term-related LB with fixed collocation and collocation. LB can be defined as a recurrent sequence of words which appears across texts in the same register and help shape distinctiveness of the register. Biber and Conrad (1999: 183) defined LB as "multi-word expressions which occur frequently and with accidental sequences of three or more words." In other words, LB is the most frequent recurring sequences of words in a given register (Biber, 2006).

Researches on LB have been carried out in various aspects and perspectives, for example the study on the relationship between LB and language proficiency (Vidacovic and Barker, 2010), LB use in academic classrooms (Kashida and Heng, 2013), comparison of LB proficiency between native and non-native English speakers (Chen and Baker, 2010; Kwon and Lee, 2014; Adel and Erman, 2012), and LB in scientific registers (Salazar, 2013; Qin, 2014, Grabowski, 2015; Pan et al., 2016).

Vidacovic and Barker (2010) study demonstrated that there is a positive relationship between LB and language proficiency. They found that the more LB students know, the higher their language proficiency level is. In addition, several studies also showed that native English speakers used types and frequencies of LB more and vary than non-native speakers (Chen and Baker, 2010; Kwon and Lee, 2014).

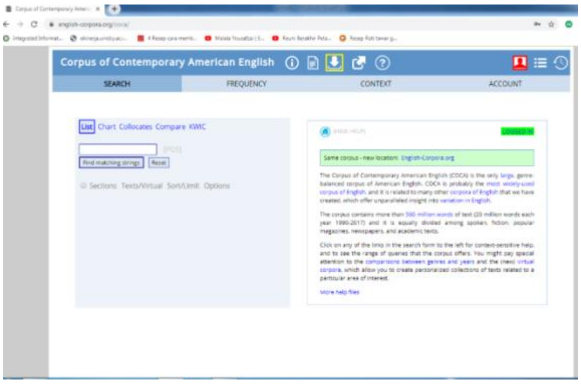
Research related to the use of LB in the scientific fields has been carried out, such as in biology (Salazar, 2013), applied linguistics (Qin, 2014), medicine (Jalali and Moini, 2014), pharmacy

(Grabowski, 2015), and telecommunication (Pan et at., 2016). In contrast, little attention has been given to the use of bundles in the register of English for specific purposes (ESP). In this gap, this paper aims to introduce the use of COCA and Lexical Tutor, and to feature Antconc as a corpus resources in analyzing LB.

III. BENEFITING FROM COCA

COCA is the largest English corpus resource availably on the internet. It is free access and serves the balanced corpus of American English. COCA was created by Mark Davies of Brigham Young University in 2008 and until recently it is used by more than ten thousands of users every month.The corpus consists of more than 500 million words which are updated every year. The corpus covers five genres. They are spoken (transcripted text of TV and radio programs), fiction (short stories and play, books and movie scripts), newspapers (wide range magazines from variety of topics such as local news, opinions, sports, financial section), popular magazines (from various magazine covering health, home and gardening, women, religion, etc.), and academic journals (more than 100 peer-reviewed journals). The COCA website is easy to access with a simple interface for learners to use with a limit number of queries per day. For full corpus texts, it is available at cost. Below is the preview of COCA website.

Figure 2. Interface of COCA


Corpus of Contemporary American English
english-corpora.org/coca/
SEARCH
FREQUENCY
CONTEXT
ACCOUNT
List Chart Collocates Compare KWIC
[POS]
Find matching strings
Reset
Sections Texts/Virtual Sort/Limit Options
LOGGED IN

Same corpus - new location: English-Corpora.org

The Corpus of Contemporary American English (COCA) is the only large, genre-balanced corpus of American English. COCA is probably the most widely-used corpus of English, and it is related to many other corpora of English that we have created, which offer unparalleled insight into variation in English.

The corpus contains more than 560 million words of text (20 million words each year 1990-2017) and it is equally divided among spoken, fiction, popular magazines, newspapers, and academic texts.

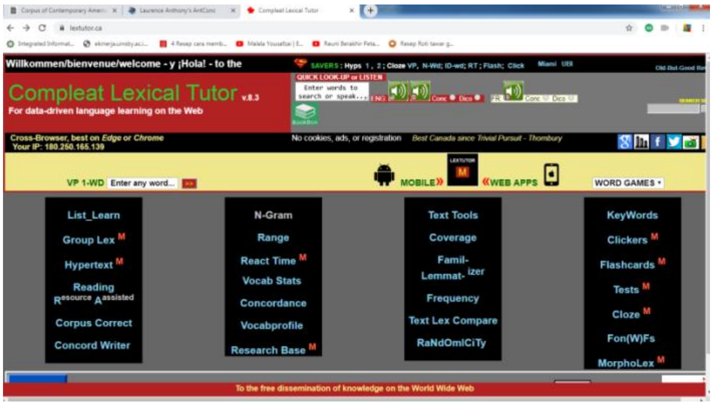
Click on any of the links in the search form to the left for context-sensitive help, and to see the range of queries that the corpus offers. You might pay special attention to the comparisons between genres and years and the (new) virtual corpora, which allow you to create personalized collections of texts related to a particular area of interest.

More help files

LB OF ACADEMIC WORDS WITH LEXICAL TUTOR

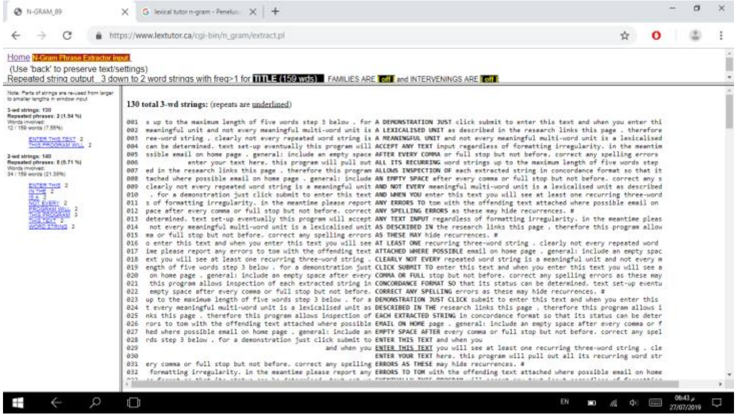
Lexical Tutor or The Compleat Lexical Tutor or shortly Lextutor is created by Tom Cobb from University of Montreal. It is aimed at data driven learning on the web which make it possible for learners to practice corpus online. The web provides driven learning for assessing their vocabulary size, knowledge of vocabulary, vocabulary in context, grammar and concordance, and researching vocabulary learning. The site offers three sections of usage, tutorial for student, for researches, and for teachers. The figure 2 below demonstrates the website screen.

Figure 1. The Compleat Lexical Tutor for data driven learning


lextutor.ca
Willkommen/bienvenue/welcome - y ¡Hola! - to the
Compleat Lexical Tutor v.8.3
For data-driven language learning on the Web
SAVERS: Hyps 1, 2; Cloze VP, N-Wd; ID-wd; RT; Flash; Click
QUICK LOOK-UP or LISTEN
Enter words to search or speak...
Cross-Browser, best on Edge or Chrome
Your IP: 180.250.165.139
No cookies, ads, or registration
Best Canada since Trivial Pursuit - Thornbury
VP 1-WD
Enter any word...
MOBILE»
«WEB APPS
WORD GAMES
List_Learn
Group Lex M
Hypertext M
Reading Resource Assisted
Corpus Correct
Concord Writer
N-Gram
Range
React Time M
Vocab Stats
Concordance
Vocabprofile
Research Base M
Text Tools
Coverage
Famil-Lemmat-izer
Frequency
Text Lex Compare
RaNdOmICiTy
KeyWords
Clickers M
Flashcards M
Tests M
Cloze M
Fon(W)Fs
MorphoLex M
To the free dissemination of knowledge on the World Wide Web

To identify LB, go to the second column and click N-Gram. Copy paste your texts into the white givencolumn. Decide the maximum string we want, ranging from 2-5 words. Then click Submit_window. The result looks like the figure 3 below.

Figure 3. Identifying Lexical Bundles using Lextutor


N-GRAM_89
lexical tutor n-gram - Penelus
https://www.lextutor.ca/cgi-bin/n_gram/extract.pl
Home
N-Gram Phrase Extractor input
EN
06:43
27/07/2019

(Use 'back' to preserve text/settings)
Repeated string output 3 down to 2 word strings with freq>1 for TITLE (159 wds) FAMILIES ARE off and INTERVENINGS ARE off

Note: Parts of strings are re-used from larger to smaller lengths in window input

3-wd strings: 130
Repeated phrases: 2 (1.54 %)
Words involved:
12 / 159 words (7.55%)
ENTER THIS TEXT 2
THIS PROGRAM WILL 2

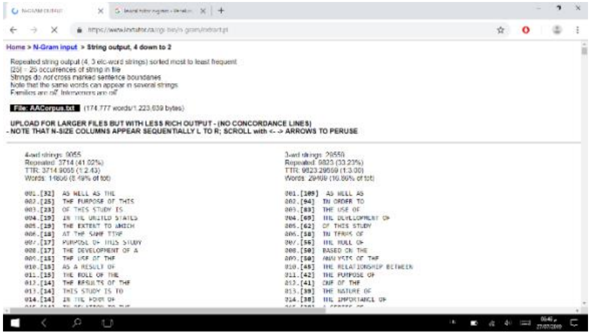
2-wd strings: 140
Repeated phrases: 8 (5.71 %)
Words involved:
34 / 159 words (21.38%)
ENTER THIS 2
IN THE 2
IS A 2
NOT EVERY 2
PROGRAM WILL 2
THIS PROGRAM 3
THIS TEXT 2
WORD STRING 2

130 total 3-wd strings: (repeats are underlined)

```
001  s up to the maximum length of five words step 3 below . for A DEMONSTRATION JUST click submit to enter this text and when you enter thi
002  meaningful unit and not every meaningful multi-word unit is A LEXICALISED UNIT as described in the research links this page . therefore
003  ree-word string . clearly not every repeated word string is A MEANINGFUL UNIT and not every meaningful multi-word unit is a lexicalised
004  can be determined. text set-up eventually this program will ACCEPT ANY TEXT input regardless of formatting irregularity. in the meantim
005  ssible email on home page . general: include an empty space AFTER EVERY COMMA or full stop but not before. correct any spelling errors
006            enter your text here. this program will pull out ALL ITS RECURRING word strings up to the maximum length of five words step
007  ed in the research links this page . therefore this program ALLOWS INSPECTION OF each extracted string in concordance format so that it
008  tached where possible email on home page . general: include AN EMPTY SPACE after every comma or full stop but not before. correct any s
009  clearly not every repeated word string is a meaningful unit AND NOT EVERY meaningful multi-word unit is a lexicalised unit as described
010   . for a demonstration just click submit to enter this text AND WHEN YOU enter this text you will see at least one recurring three-word
011  s of formatting irregularity. in the meantime please report ANY ERRORS TO tom with the offending text attached where possible email on
012  pace after every comma or full stop but not before. correct ANY SPELLING ERRORS as these may hide recurrences. #
013  determined. text set-up eventually this program will accept ANY TEXT INPUT regardless of formatting irregularity. in the meantime pleas
014   not every meaningful multi-word unit is a lexicalised unit AS DESCRIBED IN the research links this page . therefore this program allow
015  ma or full stop but not before. correct any spelling errors AS THESE MAY hide recurrences. #
016  o enter this text and when you enter this text you will see AT LEAST ONE recurring three-word string . clearly not every repeated word
017  ime please report any errors to tom with the offending text ATTACHED WHERE POSSIBLE email on home page . general: include an empty spac
018  ext you will see at least one recurring three-word string . CLEARLY NOT EVERY repeated word string is a meaningful unit and not every m
019  ength of five words step 3 below . for a demonstration just CLICK SUBMIT TO enter this text and when you enter this text you will see a
020   on home page . general: include an empty space after every COMMA OR FULL stop but not before. correct any spelling errors as these may
021   this program allows inspection of each extracted string in CONCORDANCE FORMAT SO that its status can be determined. text set-up eventu
022   empty space after every comma or full stop but not before. CORRECT ANY SPELLING errors as these may hide recurrences. #
023  up to the maximum length of five words step 3 below . for a DEMONSTRATION JUST CLICK submit to enter this text and when you enter this
024  t every meaningful multi-word unit is a lexicalised unit as DESCRIBED IN THE research links this page . therefore this program allows i
025  nks this page . therefore this program allows inspection of EACH EXTRACTED STRING in concordance format so that its status can be deter
026  rors to tom with the offending text attached where possible EMAIL ON HOME page . general: include an empty space after every comma or f
027  hed where possible email on home page . general: include an EMPTY SPACE AFTER every comma or full stop but not before. correct any spel
028  rds step 3 below . for a demonstration just click submit to ENTER THIS TEXT and when you
029                                                  and when you ENTER THIS TEXT you will see at least one recurring three-word string . cle
030                                                               ENTER YOUR TEXT here. this program will pull out all its recurring word str
031  ery comma or full stop but not before. correct any spelling ERRORS AS THESE may hide recurrences. #
032   formatting irregularity. in the meantime please report any ERRORS TO TOM with the offending text attached where possible email on home
```

Lexical Tutor also offers a collection of corpora, i.e.: Academic abstracts, Electronical engineering, Focus on vocab, Call of the wild, TC learner (student), TC learner (teacher), JPU learner, NNS-Ts in Korea, NS-Ts in Korea, French perle, and French ecrit. To use them, just choose the corpus, decide the maximum string and then click Submit_corpus, as shown in the figure 4 below.

Figure 5. Lexical Bundles Output using Lextutor



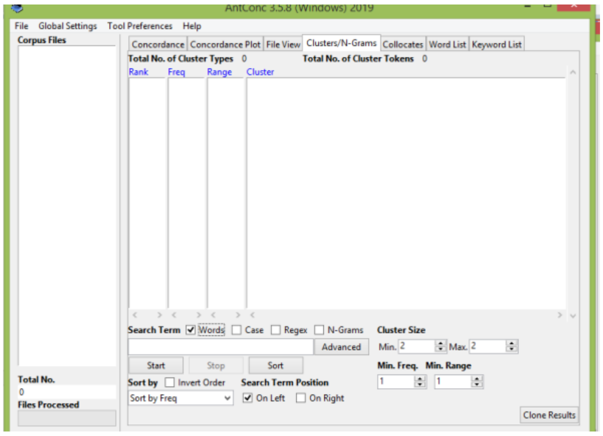
Home > N-Gram input > String output, 4 down to 2

```
Repeated string output (4, 3 etc-word strings) sorted most to least frequent
[25] = 25 occurrences of string in file
Strings do not cross marked sentence boundaries
Note that the same words can appear in several strings
Families are off. Intervenors are off

File: AACorpus.txt  (174,777 words/1,223,639 bytes)

UPLOAD FOR LARGER FILES BUT WITH LESS RICH OUTPUT - (NO CONCORDANCE LINES)
- NOTE THAT N-SIZE COLUMNS APPEAR SEQUENTIALLY L TO R; SCROLL with <- -> ARROWS TO PERUSE

4-wd strings: 9055                      3-wd strings: 29550
Repeated: 3714 (41.02%)                 Repeated: 9823 (33.23%)
TTR: 3714:9055 (1:2.43)                 TTR: 9823:29559 (1:3.00)
Words: 14856 (8.49% of tot)             Words: 29409 (16.86% of tot)

001.[32]  AS WELL AS THE                001.[109]  AS WELL AS
002.[25]  THE PURPOSE OF THIS           002.[94]  IN ORDER TO
003.[23]  OF THIS STUDY IS              003.[83]  THE USE OF
004.[19]  IN THE UNITED STATES          004.[69]  THE DEVELOPMENT OF
005.[19]  THE EXTENT TO WHICH           005.[62]  OF THIS STUDY
006.[18]  AT THE SAME TIME              006.[58]  IN TERMS OF
007.[17]  PURPOSE OF THIS STUDY         007.[56]  THE ROLE OF
008.[17]  THE DEVELOPMENT OF A          008.[50]  BASED ON THE
009.[15]  THE USE OF THE                009.[50]  ANALYSIS OF THE
010.[15]  AS A RESULT OF                010.[45]  THE RELATIONSHIP BETWEEN
011.[15]  THE ROLE OF THE               011.[42]  THE PURPOSE OF
012.[14]  THE RESULTS OF THE            012.[41]  ONE OF THE
013.[14]  THIS STUDY IS TO              013.[39]  THE NATURE OF
014.[14]  IN THE FORM OF                014.[38]  THE IMPORTANCE OF
```

IDENTIFYING LB USING ANTCONC

AntConc is a freeware, multiplatform tool for carrying out corpus linguistics research and data-driven learning (Anthony, 2014). This program has been designed not only for researchers but also specifically for teachers and learners in aclassroom context. This freeware is ideal for those with limited budget but it serves high quality and complex tasks for corpus analysis. Although it is a freeware, AntConc a powerful concordance, tool for analysing cluster and LB, word frequency generator, and user friendly with intuitive graphical interface.

Figure 6. AntConc



Total No. of Cluster Types 0 Total No. of Cluster Tokens 0

Search Term ☑ Words □ Case □ Regex □ N-Grams

Cluster Size Min. 2 Max. 2

Min. Freq. 1 Min. Range 1

Sort by □ Invert Order Sort by Freq

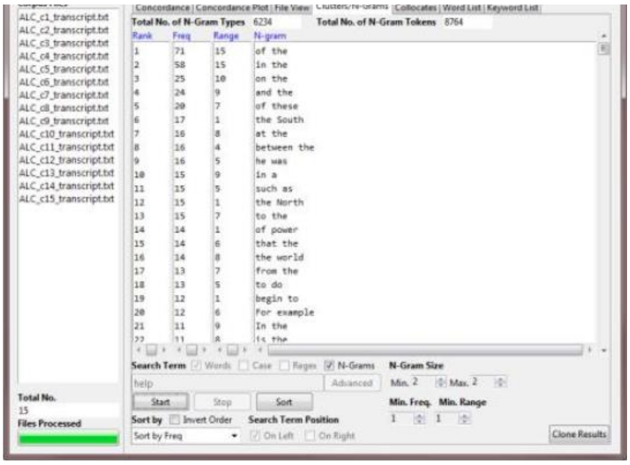
Search Term Position ☑ On Left □ On Right

Total No. 0

Files Processed

Learners of English as FL more often find difficulty in understanding collocation and word clusters. As a result, they need to be given more chance to work with texts in a specific technical or scientific field to enable them better understand the words as the lexical unit is very often longer than a single word occurs. AntConc provides a tool to investigate multi-word units using Cluster/N-Grams.The tool shows clusters of words or LB and order them alphabetically. The search term can be determined with cluster size as minimum and maximum words.It is also possible to specify a minimum frequency threshold for theclusters generated. The figure below displays the use of N-Gram or LB of AntConc.

Figure 4. Lexical Bundles list using AntConc



ALC_c1_transcript.txt
ALC_c2_transcript.txt
ALC_c3_transcript.txt
ALC_c4_transcript.txt
ALC_c5_transcript.txt
ALC_c6_transcript.txt
ALC_c7_transcript.txt
ALC_c8_transcript.txt
ALC_c9_transcript.txt
ALC_c10_transcript.txt
ALC_c11_transcript.txt
ALC_c12_transcript.txt
ALC_c13_transcript.txt
ALC_c14_transcript.txt
ALC_c15_transcript.txt

Total No. of N-Gram Types 6234 Total No. of N-Gram Tokens 8764

Rank	Freq	Range	N-gram
1	71	15	of the
2	58	15	in the
3	25	10	on the
4	24	9	and the
5	20	7	of these
6	17	1	the South
7	16	8	at the
8	16	4	between the
9	16	5	he was
10	15	9	in a
11	15	5	such as
12	15	1	the North
13	15	7	to the
14	14	1	of power
15	14	6	that the
16	14	8	the world
17	13	7	from the
18	13	5	to do
19	12	1	begin to
20	12	6	For example
21	11	9	In the
22	11	8	is the

Search Term ☑ Words □ Case □ Regex ☑ N-Grams

help

N-Gram Size Min. 2 Max. 2

Min. Freq. 1 Min. Range 1

Sort by □ Invert Order Sort by Freq

Search Term Position ☑ On Left □ On Right

Total No. 15

Files Processed

and demonstrate the main menus of this tool.
1) Click on the N-Grams option above the search entry box.
2) Choose the appropriate ordering options.
3) Press the Start button.
4) Click on the n-gram to generate a set of KWIC lines using the text as the search term.
5) Click on the Clone Results button to create a copy of the results so that different sets of results can be compared.

IV. CONCLUSION

This paper has discussed the relevance of corpus resources and tools for analysing lexical bundles. It has demonstrated evidence that corpus-driven tools can present precious insights for vocabulary enrichment in English learning. Resources such as COCA, The Compleat Lexical Tutor and AntConc provide a better understanding of the different and unusual nature of English. By data driven learning, the tools may serve to maximize the benefits from English vocabulary learning into the language classroom.

REFERENCES

[1] Anthony, L. (2019). Antconc. Retrieved from Anthony, L. (2019). AntConc (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

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

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

[4] Grabowski, L. (2015). Keywords and lexical bundles within English pharmaceutical discourse: A corpus-driven description. *English for Specific Purposes, 38,* 23–33.

[5] Granger, S., &Meunier, F. (2008). *Phraseology in language learning and teaching. Where to from here?* In Meunier, F. & Granger, S. (Eds.), Phraseology in foreign language learning and teaching. John Benjamins.

[6] Jalali, Z. S., &Moini, M. R. (2014). Structure of lexical bundles in introduction section of medical research articles. *Procedia Social and Behavioral Sciences*, *98*, 719–726.

[7] Kashiha, H., &Heng, C. S. (2013). An Exploration of Lexical Bundles in Academic Lectures: Examples from Hard and Soft Sciences. *The Journal of Asia TEFL*, *10*(4), 133–161.

[8] Kwon, Y.-E., & Lee, E.-J. (2014). Lexical Bundles in the Korean EFL Teacher Talk Corpus: A Comparison Between Non-native and Native English Teachers. *The Journal of Asia TEFL*, *11*(3), 73–103.

[9] Pan, F., Reppen, R., &Biber, D. (2016). Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in telecommunications research journals. *Journal of English for Academic Purposes*, *21*, 60–71.

[10] Pawley, A., &Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (eds). *Language and communication*, 191-225

[11] Qin, J. (2014). Use of formulaic bundles by nonnative English graduate writers and published authors in applied linguistics. *System*, *42*(1), 220–231.

[12] Salazar, D. J. L. (2013). *Biomedical English: A corpus based approach*. Oxford: John Benjamins.

[13] Schmitt, N., & Carter, R. (2004). *Formulaic sequences in action: An introduction*. In N. Schmitt (Ed.), Formulaic sequences: Acquisition, processing, and use. Philadelphia: John Benjamins.

[14] Scott, M. (1997). PC analysis of key words and key key words. *System*, *25*(1), 233-245.

[15] Stubbs, M (2005). The most natural thing in the world: quantitative data analysis on multi-word sequences in English. Paper presented at *Phraseology* 2005, 13–15 October 2005, Louvainla-Neuve.

[16] Vidakovic, I., & Barker, F. (2010). Use of words and multi-word units in skills for life writing examinations. *Cambridge ESOL: Research Notes*, *41*, 7–14.

[17] Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford: Oxford University Press.