Data Driven Learning by Discovering Lexical Bundles Using Corpus Resources

Kamal Yusuf
Faculty of Adab and Humanities
UIN Sunan Ampel 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 (Stubbis, 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), medicine (Jalali and Moini, 2014), pharmacy
(Grabowski, 2015), and telecommunication (Pan et al., 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 available 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 (transcribed text of TV and radio programs), fiction (short stories and play, books and movie scripts), newspapers (wide range of 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**

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**

To identify LB, go to the second column and click N-Gram. Copy paste your texts into the white given column. 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**

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.
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 a classroom 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.

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 the clusters generated. The figure below displays the use of N-Gram or LB of AntConc.

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


