AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit

Laurence ANTHONY
Center for English Language Education in Science and Engineering
School of Science and Engineering
Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
ant_web@antlab.sci.waseda.ac.jp

Abstract

AntConc is a freeware, multi-platform, multi-purpose corpus analysis toolkit, designed specifically for use in the classroom. It hosts a comprehensive set of tools including a powerful concordancer, word and keyword frequency generators, tools for cluster and lexical bundle analysis, and a word distribution plot. In this paper, I will describe each of these tools, and explain their value to learners. Then, I will discuss the current limitations of the software, before explaining how I hope it will be improved in future releases.

Keywords: *corpus linguistics, concordancer, collocation, software, educational technology*

1 Introduction

Over the past ten years, corpora of language data have started to play an increasingly important role in determining how languages are taught (Coniam, 2004). As Hunston (2002) writes, corpora have been applied in a wide range of areas, including translation studies, stylistics, and grammar and dictionary development. In the classroom, one of their most important applications has been as part of a data driven approach to learning (Johns, 1997). The effectiveness of this is highlighted in Noguchi's (2002) study, where she describes how graduate students in science and engineering fields are able to improve their writing skills through the analysis of small sized corpora from their target fields.

A corpus is virtually useless without some kind of computer software tool to process it and display results in an understandable way. Two of the most popular software tools for this are *MonoConc Pro*[1], and *WordSmith Tools*[2], although many other concordancers and corpus analysis programs have also been developed. Strangely, few of these programs have been designed specifically for learners in a classroom context. Rather, they have tended to be aimed at researchers, and thus either include a wide range of features rarely needed by most learners (for example *WordSmith Tools*), or a very limited number of features sufficient to perform only a specific task (for example *Web Concordancer*[3]). In addition, the design of the graphical user interface (GUI) has been less of a concern in many cases, resulting in overly complex or rudimentary interfaces that lack the familiar feel of a modern windows based application.

In this paper, I will describe *AntConc*[4], a corpus analysis toolkit designed specifically for use in the classroom. AntConc is a freeware application, making it ideal for individuals, schools or colleges with a limited budget, and runs on both Windows and Linux/Unix based systems. Although it has a freeware license, it includes an easy-to-use, intuitive graphical user interface and offers a powerful concordancer, word and keyword frequency generators, tools for cluster and lexical bundle analysis, and a word distribution plot. In the following sections, I will describe each of the tools in the latest edition of the program (version 3.0), and explain how they can be of value to learners. I will then detail the program's limitations, before concluding the paper with a discussion on its future development.

2 Concordancer Tool

The central tool used in most corpus analysis software, including AntConc, is the concordancer. As Sun & Wang (2003) describe, concordancers have been shown to be an effective aid in the acquisition of a second or foreign language, facilitating the learning of vocabulary, collocations, grammar and writing styles.

Figure 1 shows a screenshot of AntConc while a user is operating the Concordancer Tool. As with all other tools in the program, the Concordancer Tool is designed so that the most common operations are accessible directly on the main screen. Lonfils & VanParys (2001) explain that this is an essential feature of good software design as it avoids the need for confusing pull-down menus and additional windows. It can also be seen that the tool's interface widgets, such as check buttons, lists, and window adjusters, have the native look and feel of the operating system, and the same functionality as standard software applications.

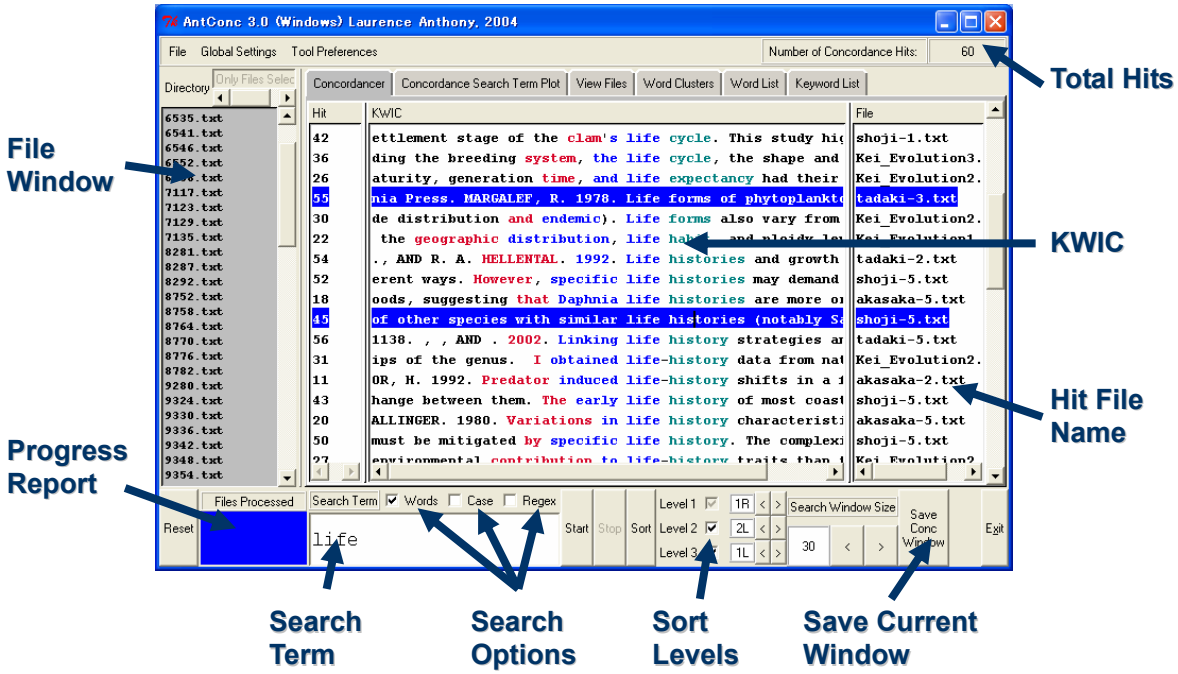


Figure 1 KWIC Concordancer Tool

The Concordancer Tool has a wide range of features that make it an extremely effective tool not only for learners, but also teachers and researchers. These are summarized as follows:

1. Search terms can be substrings, words, or phrases, and can be case sensitive or insensitive. They can be embedded with a wide range of wildcards, which the user can assign to any particular character or string of characters via a menu option.

2. Search terms can be defined as full regular expressions (REGEX), offering the user access to extremely powerful and complex searches.

3. Three levels of sorting of KWIC (Key Word In Context) lines are possible, with user definable highlight colors at each level.

4. If a user clicks on any search term in the KWIC results display, the program will automatically open the View Files tool (described later) and show the search term hit in the original data file.

5. The KWIC results display is divided into columns, in which the hit number, KWIC line, and file name are shown separately. As in all other tools, each column can be either displayed or hidden, and standard selection methods can be used to save data in the columns or rows to the clipboard or a text file.

3 Concordance Search Term Plot Tool

The main purpose of the Concordancer Tool is to show *how* a search term is used in a target corpus. For users who want to see *where* a search term appears AntConc offers the Concordancer Search Term Plot Tool, shown in Figure 2. Here, each box represents a file in which multiple lines represent the relative positions at which search term hits can be found. From this display, it is easy to see not only how often a search term appears in a corpus of data, but also where and in what distribution. This can be an effective aid, for example, in determining where particular phrases such as "we" or "in this paper" are used in a research article.

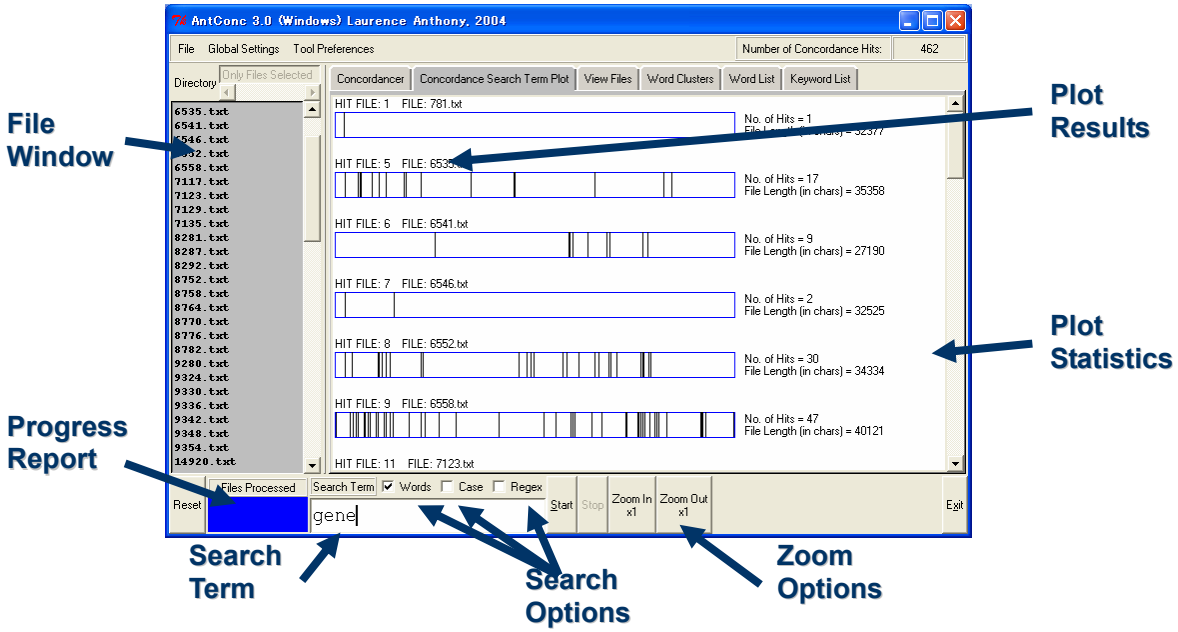


Figure 2 Concordance Search Term Plot Tool

4 View Files Tool

The View Files Tool of AntConc is shown in Figure 3. As mentioned above, when a user clicks on a search term in the results display of the Concordancer Tool, the View Files tool is used to display the search term in the original file. However, the View Files Tool can be used independently to search for any substring, word, phrase or regular expression in a target file, offering the user a very powerful text search engine. As can be seen in the figure, all hits are displayed in a highlight color, and buttons and keyboard shortcuts can be used to jump to a specified hit anywhere in the file.

5 Word List / Keyword List Tools

One of the first things that a user will do when analyzing a new corpus is to generate a list of all the words in the corpus. Word lists are useful for highlighting interesting areas in a corpus and suggesting problem areas. As Bowker & Pearson (2002) note, they can also be used to find the lemmas of words in a corpus, or

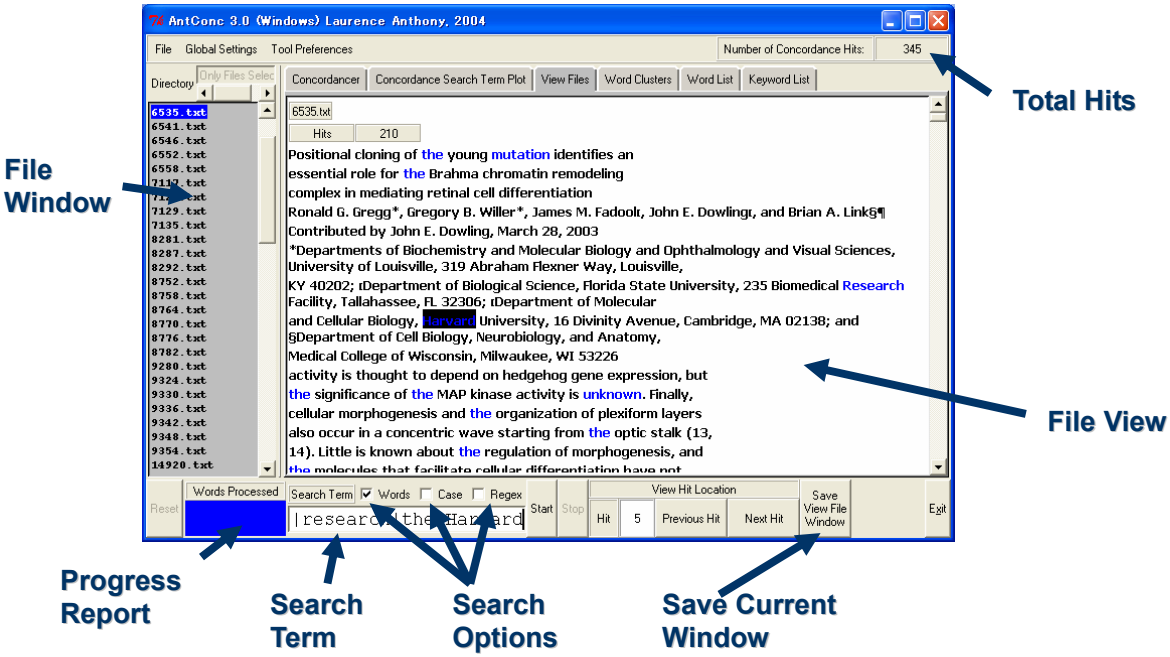
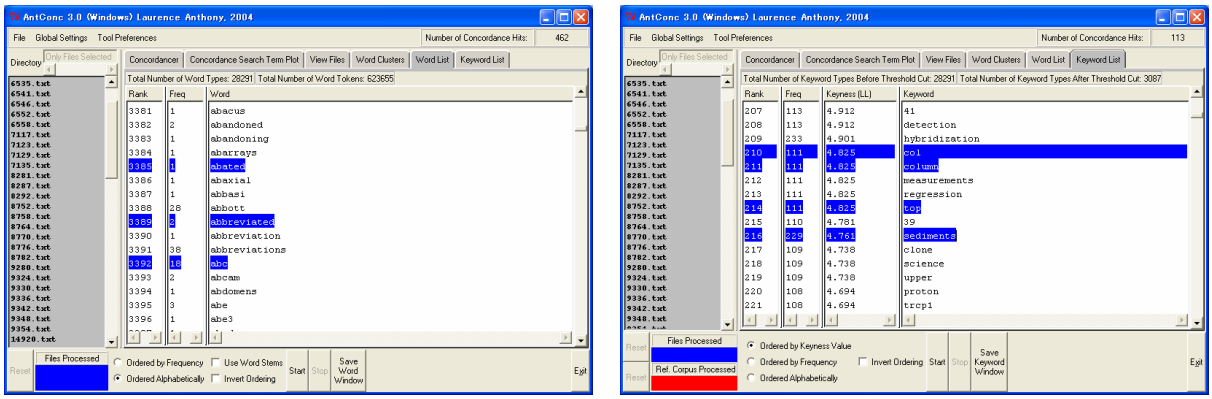


Figure 3 View Files Tool



a) Word List Tool b) KeyWord List Tool

Figure 4 Word List and Keyword List Tools

families of related word forms. The Word List Tool of AntConc is shown in Figure 4a.

Hockey (2001) states that an ideal word list generation program should be able to sort words into alphabetical or frequency order. The Word List Tool of AntConc offers these features and also the ability to count words based on their 'stem' forms. In order to avoid counting high frequency functional words when generating a word list, a stop list can be specified in the Word List Tool's preference window either by direct input from the keyboard or from a separate file. In addition, users can specify the reverse of a stop list, i.e., a list of only the words that should be counted.

As experienced users of corpus analysis tools will know, word lists usually tell us little about how *important* a word is in a corpus. Therefore, AntConc offers a Keyword List Tool, which finds which words appear unusually frequently in a corpus compared with the same words in a reference corpus that is also specified by the user. The Keywords Tool operates in an almost identical way to the

KeyWords tool in WordSmith Tools, calculating the 'keyness' of words using either the chi-squared or log likelihood statistical measures (Kilgarriff, 2001), and offering the user the option of displaying or hiding unusually infrequent keywords (or negative-keywords) in the preferences window. The Keyword List Tool is shown in Figure 4b.

6 Word Clusters / Bundles Tool

Research has shown that collocations and other multi-word units such as phrasal verbs, and idioms are particularly difficult for learners to acquire (Nesselhauf & Tschichold, 2002). Their importance is even greater if the learner is working with texts in a highly technical or scientific field, as the lexical unit is very often longer than a single word (Bowker & Pearson, 2002).

In AntConc, multi-word units can be investigated using the Word Clusters Tool, shown in Figure 5. This tool displays clusters of words that surround a search term and orders them alphabetically or by frequency. The search term can be specified as a substring, word, phrase or regular expression as in the Concordancer, Plot and View File tools, and the number of additional words to the left and right of the search term can also be specified. It is also possible to set a minimum frequency threshold for the clusters generated.

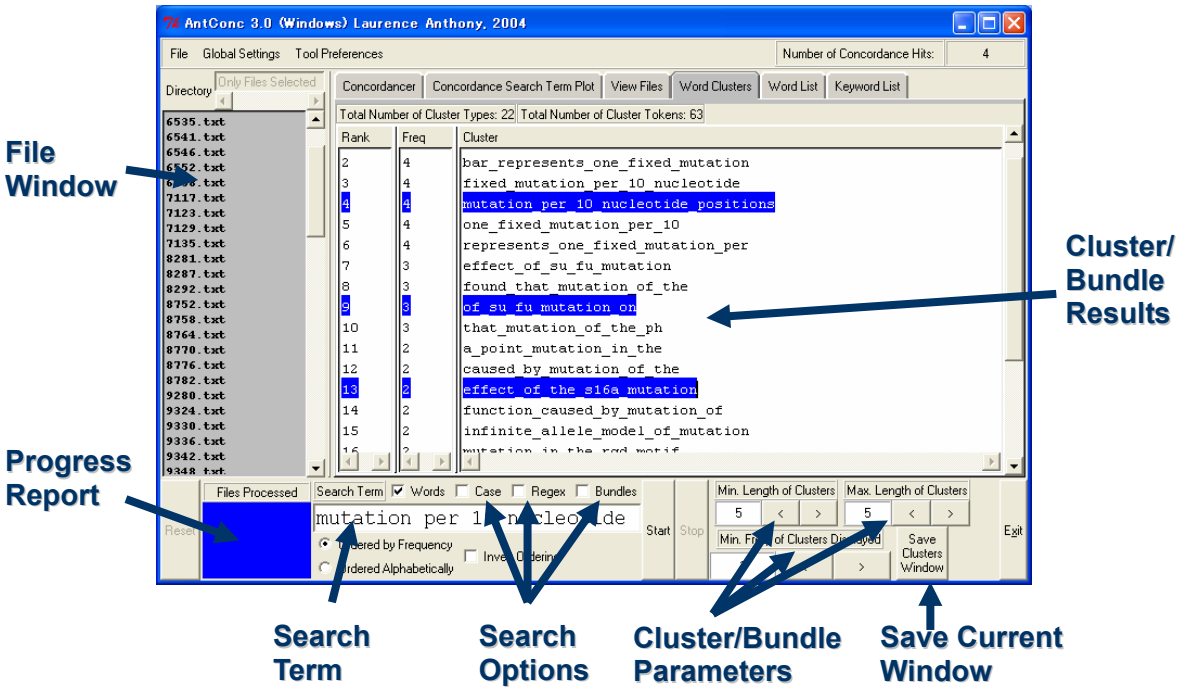


Figure 5 Word Clusters / Bundles Tool

An alternative way to search for multi-word units is to find lexical bundles (Biber et al., 1999), which are equivalent to n-grams, where n can vary usually between two and five words. Few corpus analysis programs offer this feature (Coniam, 2004), but AntConc includes lexical bundle searches as an option in the Word Clusters Tool.

7 Limitations of AntConc

AncConc performs all operations directly on the raw texts of the corpus. This is useful in that the user is often switching or modifying the target corpus for a particular need, as the program does not need to do any pre-processing of the data, for example, creating an index (Hockey, 2001). On the other hand, because AntConc does not use an index, it can only work effectively with small scale corpora. Nevertheless, as McEnery & Wilson (2001) note, one of the major trends in corpus linguistics over the

past few years is the increased interest in very small, highly specialized corpora. Small corpora can be used for a great many different purposes, as exemplified by Ghadessy et al. (1996) and Noguchi (2004).

Most corpus analysis programs offer users the ability to see the collocates of a search term in a table, where the frequency of the most common words to the left or right of the search term are indicated. Learners often find such tables difficult to interpret and so the current version of AntConc offers no implementation of this feature. However, for advanced learners this can be a severe disadvantage that will be addressed in the next release of the program.

Some programs offer detailed statistics related to the corpus and search results. Again, it was felt that these would overwhelm many learners and so the advice given by Hockey (2001) was followed. Namely, that the program should not include such statistics but instead offer an easy way to copy and paste results into a spreadsheet program for analysis later. As described earlier, the results in all display windows of AntConc can easily be copied and pasted directly into a spreadsheet program using simple keyboard shortcuts.

One of the weakest areas of AntConc is in the handling of annotated data such as data encoded in HTML/XML format. Although AntConc offers a simple way to view or hide embedded tags used in HTML/XML and other annotation methods, much more sophisticated methods need to be implemented if the full power of annotated data is to be realized.

8 Conclusions and Future Developments

AntConc is a lightweight, simple, and easy to use corpus analysis toolkit that has been shown to be extremely effective in a classroom context (Noguchi, 2004). Although it does not include all the tools and features offered in the most popular commercial applications, it includes many of the essential tools needed for the analysis of corpora, with the added benefit of an intuitive interface, and a freeware license.

To date, there have been 19 releases of the program since its launch in 2002, including three major upgrades. There are also plans to release a new version of the software in the near future that addresses some of the limitations described in the previous section. The first improvement will be a redesign of the View Files Tool making it operate with far greater speed. The current tool is able to handle files with ambiguous line endings but this comes with a heavy loss in speed. The next release will also include a tool to view collocates, and the ability to sort word lists alphabetically from both the beginning and end of words; a feature recommended by Hockey (2001).

In a later release, it is hoped that AntConc will be improved to handle annotated data, in particular XML, in a much more powerful and intuitive way. XML data includes header definitions that if extracted, can be used as part of search criteria. If this extraction can be carried out automatically by the software, it will enable users to access these definitions without any knowledge of the annotation method.

Finally, a detailed user manual and accompanying tutorial video are planned for the software, where the operation of each tool will be explained with concrete examples and a step-by-step guide.

Acknowledgements

This research was supported by a Grant-in-aid for Scientific Research by the Japan Society for the Promotion of Education, Science, Sports and Culture, Japan (No. 16700573), and by a Waseda University Grant for Special Research Projects, Japan (No. 2004B-861).

Notes

1. Information and download instructions available at: http://www.monoconc.com/
2. Information and download instructions available at: http://www.lexically.net/wordsmith/

3. Information and download instructions available at:
p://vlc.polyu.edu.hk/concordance/aboutweb.htm
4. Information and download instructions available at: http://www.antlab.sci.waseda.ac.jp/

References

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. 1999. *Longman grammar of spoken and written English*. London: Longman.
Bowker, L. and Pearson, J. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London/New York: Routledge.
Coniam, David. 2004. Concordancing oneself: Constructing individual textual profiles. *International Journal of Corpus Linguistics*, 9(2), 271–298.
Ghadessy, M., A. Henry and R. L. Roseberry. 1996. *Small Corpus Studies and ELT: theory and practice*. Amsterdam: John Benjamins.
Hockey, S. 2001. Concordance Programs for Corpus Linguistics. In Rita C. Simpson and John M. Swales *(eds.), Corpus Linguistics in North America: Selections from the 1999 Symposium*, Ann Arbor: University of Michigan Press, pp. 76-97.
Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press
Johns, T. 1997. Contexts: the Background, Development and Trialling of a Concordance-based CALL Program. In A. Wichmann, S. Fligelstone, T. McEnery, and G. Knowles (eds.), *Teaching and Language Corpora*. London: Longman, pp. 100-115.
Kilgarriff, Adam. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1), 97-133.
Lonfils, C. and Vanparys, J. 2001. How to design user-friendly CALL interfaces. *Computer Assisted Language Learning*, 14(5), 405-417.
McEnery, T. and Wilson, A. 2001. *Corpus Linguistics. An Introduction.* Second edition. Edinburgh: Edinburgh University Press.
Nesselhauf, N. and Tschichold, C. 2002 Collocations in CALL. An investigation of vocabulary-building software for EFL. *Computer Assisted Language Learning*, 15(3), 251-279.
Noguchi, J. 2004. A genre analysis and mini-corpora approach to support professional writing by nonnative English speakers. *English Corpus Studies*, 11, 101-110.
Sun, Y. C. & Wang, L. Y. 2003. Concordancers in the EFL Classroom: Cognitive Approaches and Collocation Difficulty. *Computer Assisted Language Learning*, 16 (1), p. 83-94.