Developing a Freeware, Multiplatform Corpus Analysis Toolkit for the Technical Writing Classroom

Tutorial

—Feature by
LAURENCE ANTHONY

***Abstract*—*This paper describes the development of the AntConc corpus analysis toolkit, originally designed for use in a technical writing course at Osaka University, Japan, but now adopted in institutions throughout the world as an easy-to-use, freeware, multiplatform alternative to the many commercial concordance programs. First, I will explain how the software was originally tailored to the needs of students in the Osaka writing course and later to a general audience through the requests and feedback from teachers and students around the world. Then, I will give an overview of tools in the most recent version of AntConc and explain their value using examples from the classroom. Finally, I will discuss some of the software's limitations and future developments, and suggest applications in professional communication.*

***Index Terms*—*Corpus linguistics, educational technology, software, technical writing.*

Since the early 1990s, corpora have played an increasingly important role in determining how languages are taught [1]. As Chapelle describes, there appears to be a "corpus revolution" under way [2, p. 38]. Corpora are now being applied in a wide range of areas, including translation studies, stylistics, and grammar and dictionary development [3]. In the classroom, Johns has established that learners can use corpora, in a so-called data-driven approach to learning, to investigate for themselves the way language is used in target contexts [4]. This practice can be particularly effective in the technical writing classroom, where learners are often from different fields, each with its own set of characteristic language features [5]. As Levis and Levis discuss, it is almost impossible for the teacher to know all these features and deal with them on a one-to-one basis [6].

A further advantage of data-driven learning with corpora is that it helps to develop skills in corpus analysis that will serve learners throughout their professional lives [2]. For example, most scientists and engineers who are using English as a nonnative language in the workplace do not have coworkers who are nonnative speakers to consult with on technical communication matters. Thus, when a language problem is confronted they will inevitably turn to standard textbooks, style guides, and dictionaries to find the answer, often with little success. With suitable software and a narrowly defined corpus of target texts, on the other hand, clear and comprehensive answers to many questions can be found rapidly. In an interesting parallel, a recent trend among scientists and engineers is to search the world wide web to resolve language issues, which is in effect using the web itself as a corpus.

Just as the web is virtually useless without a web browser, a language corpus is virtually useless without a software tool (e.g., a concordancer) to process it and display results in an easy to understand way. Although many corpus analysis programs now exist, including WordSmith Tools (http://www.lexically.net/wordsmith/), MonoConc Pro (http://www.monoconc.com/), WordPilot (http://home.ust.hk/~autolang/whatis_WP.htm), and Web Concordancer (http://vlc.polyu.edu.hk/concordance/aboutweb.htm), few have been developed specifically for learners in a classroom context. Rather, they are usually aimed at researchers in applied linguistics and thus either include a wide range of features rarely needed by most learners or a very limited number of features, sufficient to perform only specific tasks. In addition, the design of the graphical user interface (GUI) has been less of a concern in most cases, resulting in overly complex or rudimentary interfaces lacking the familiar feel of a modern windows-based application.

In this paper, I will describe the development of AntConc (http://www.antlab.sci.waseda.ac.jp/), a corpus analysis toolkit designed by the author for classroom use that includes a powerful concordancer, word and keyword frequency generators, tools for cluster and lexical bundle analysis, and a word distribution plot. AntConc is a freeware, multiplatform application, making it suitable for

Manuscript received December 15, 2005; revised March 3, 2006.
The author is with the Center for English Language Education in Science and Engineering, Faculty of Science and Engineering, Waseda University, Tokyo 169-8555, Japan (email: anthony@antlab.sci.waseda.ac.jp).

IEEE DOI 10.1109/TPC.2006.880753

0361-1434/$20.00 © 2006 IEEE

individuals, schools, or colleges with a limited budget who are running either Windows or Linux/Unix systems. Although designed for the classroom, this does not mean that the software cannot be used in other contexts. On the contrary, the program's ease-of-use and flexibility have made it popular among researchers and workplace scientists and engineers. Distributed as a single executable file, it can simply be copied onto the user's computer and launched without any installation.

In the following section, I will describe the background of AntConc, explaining how the software was originally tailored to the needs of Japanese technical writing students and later to a more general audience through the requests and feedback from users around the world. I will also summarize the program's main features. In the section titled Getting Started with AntConc, I will briefly explain what users need to do before they can start working with AntConc. Then, in subsequent sections, I will give an overview of the GUI and tools in AntConc and explain their relevance in the technical writing classroom using a number of real-world writing problems. I should emphasize, however, that these are only illustrative and that the range of problems that AntConc can be applied to is much more extensive. For a detailed introduction to the various techniques used in corpus linguistics, almost all of which can be practiced using AntConc, the reader is encouraged to read McEnery and Wilson [7], Biber et al. [8], Kennedy [9], Meyer [10], and Hunston [3]. In the section titled Limitations and Future Developments, I will discuss some of the current limitations and future developments of AntConc. In the final section, I will suggest some uses of the software for professional communicators in educational institutions and the workplace, and I will make my conclusions.

BACKGROUND

In 2002, Osaka University, in the west of Japan, was planning a new technical writing course for its 700 graduate school engineering students. As the students would be coming from diverse disciplines (e.g., mechanical engineering, biological sciences, and computer science), it was decided that the class should focus on observing target texts in the students' individual fields, analyzing and classifying these texts, hypothesizing about field-specific writing principles, and finally applying what was learnt in the writing of original texts [11]. A key goal of the course, therefore, was for students to create their own mini-corpora of target texts, which would then be used as part of in-class study. Training in corpus linguistics was made an essential part of the course as it would enable students to identify and hypothesize on the important and unique language features in their specialized corpora, following Johns' data-driven approach to learning [4]. For full details on the background and overall aims of the course, see Noguchi [11].

The organizers at Osaka faced two problems. First, the language laboratory in which the classes would be held was a Linux environment, preventing the students from using any of the standard Windows-based corpus analysis software applications. Second, the organizers had little funding for the development of new, in-house corpus linguistics software, designed for Linux and geared toward students who might not be familiar with computers.

Around this time, my own experience and frustrations with commercial software had led me to start developing a simple corpus analysis program that was intuitive to use and had the look and feel of a modern windows-based application. Using the Perl programming language, a prototype of the program was written in a Windows environment and a beta-version released as freeware to research colleagues. Through our work on various other projects, one of the Osaka course organizers soon became aware of my program and asked if I would be prepared to modify it for the Osaka course after porting it to Linux.

The advantages for both parties were obvious. The Osaka course organizers would obtain a freeware corpus analysis program designed specifically for technical writing students that would operate within the technical limits of their language laboratory. For me, the course's teachers and students would provide a valuable source of feedback, showing which of my program's features and tools were performing well and pointing out problem areas, such as bugs and GUI design issues. I was also able to gain a great deal of motivation from seeing so many students in the classroom using the software to solve real-world writing problems, a true solace in the lonely world of software programming.

Through the first part of 2002, the software, now named AntConc, was ported to Linux and tested. Then, in September 2002, the first Linux version of AntConc was uploaded onto the Osaka systems in time for the start of the new course. The software was an immediate success, providing the ideal tool for teaching corpus linguistics. Although it is beyond the scope of this paper to describe all of the activities in which the program has since been used, several examples are given in the following sections.

After the release of the Linux version of AntConc, both the Linux and Windows versions of AntConc were uploaded onto my website so that any teacher or student around the world could download the software free of charge for nonprofit use. This generated wide interest in the program, especially after it was included

in Morphix NLP (http://morphix-nlp.berlios.de/), a CD Linux distribution containing a wide range of natural language processing (NLP) tools. Since 2002, AntConc has been downloaded over 20,000 times, and downloads of Morphix NLP have topped 10,000. As both programs are distributed free of charge and can be installed on entire sites without an additional license, the actual number of AntConc users is unknown, but it is clearly well in excess of the above figures. For example, in the Osaka writing course alone over 2,000 students have had access to AntConc, and there are now plans to install the software on all public-access computers at Waseda University in Japan, with a student base of over 50,000.

Since 2002, I have continued to work with the Osaka staff to update the program. However, I have also been fortunate to receive reports from many other users around the world, including those in the US, UK, Spain, Germany, and the Netherlands, who have described their successes with AntConc and mentioned features they would like to see added to it. Provided that the requests do not go against my original premise that AntConc is designed primarily for technical writing students and should be intuitive and easy to use, I have tried to implement as many of these changes as possible. Examples of such improvements include adding multiple level sorting of results; the ability to change the style, size, and color of fonts; support for html/xml file formats; wildcards; and the ability to show or hide long file names and embedded tag information.

From its rather simple origins, AntConc has now undergone 16 minor and three major upgrades. As of writing, the latest version is AntConc 3.0.1, released in March 2005. A summary of the tools and features in this version can be found in Table I.

Getting Started With AntConc

To maintain the security of their computer systems, Osaka University, like most other universities, does not allow individual users to install software programs. Rather, users make formal requests to their departments, which then make formal requests to the computer administration technical staff. If the request is accepted, the new software is eventually installed on the necessary machines, usually by technical staff. Often the process can become very complex and time consuming and, in many cases, has to be repeated for each update of the software in question. Unfortunately, most of the popular commercial corpus linguistics programs require installation, making system management a serious concern for many institutions.

To overcome this problem, AntConc was designed to be distributed as a single, stand-alone, executable (exe) file, requiring no installation and operating solely in a computer's active (RAM) memory. For these reasons, AntConc can be rapidly introduced or updated on classroom computers and so avoids all the above system management issues. In fact, the software does not even need to be copied to the target computer and instead can be carried on a portable device such as a CD-ROM or USB flash memory device. Also, as there is only one file to deal with, starting the program is a simple matter of double-clicking on the program icon (see Fig. 1).

In addition to AntConc itself, the only other thing a user requires is a prepared corpus. As mentioned earlier, at the beginning of the writing course at Osaka University, students are asked to create their own mini-corpora by downloading or scanning texts from their target disciplines into a text editor, such as Microsoft WordPad, and saving these in a folder [11]. These texts are sometimes full research articles, but students are also encouraged to create corpora of partial texts, such as research articles' titles or abstracts (see sections titled Concordancing with the AntConc Concordancer Tool and Finding Positional Information with the AntConc Concordance Search Term Plot Tool). Usually, these files are saved as plain text (txt) files, although, as shown in Table I, AntConc can also handle data in html or xml format.

The AntConc GUI

The GUI of AntConc was developed using the Perl/Tk toolkit, which enables the program to adopt the

TABLE I
SUMMARY OF TOOLS AND FEATURES IN ANTCONC 3.0.1

Tool and Features
Freeware license
Small memory requirement (~2MB)
Multiplatform • Windows 95 or later • Unix/Linux
Extensive set of text analysis tools • KWIC concordance • Search term distribution plot • Original file view • Word clusters/lexical bundles • Word lists • Keyword lists
Powerful search features • Regular expressions (REGEX) • Extensive wildcards
Multiple-level sorting
HTML/XML tag handling
Unicode support
Easy-to-use, intuitive GUI

native feel and look of the target operating system (see Fig. 1). For students at Osaka University lacking advanced computer skills, this is an important feature as it allows them to rapidly learn the basic program operations without instruction manuals or help pages. As Lonfils and Vanparys explain, this ease-of-use comes about because the program matches the user's habits and expectations [12]. For example, selecting a set of files for processing in AntConc is synonymous with selecting files to open in Microsoft Word. Similarly, other common procedures, such as saving results, copying and pasting, and selecting options are identical to their Microsoft Word counterparts.

To further ease the use of AntConc, all the program's tools and major features can be accessed directly from the top window. This avoids the need for a multitude of confusing pull-down menus or subwindows, which can be especially difficult for students to navigate in a teacher-led classroom [12]. Also, all tools follow a consistent design format, so that mastery of one will naturally lead to the mastery of all.

One further feature of the program, which was added at the request of a teacher in the Netherlands, is the ability to change the size, style, and color of fonts. This may seem a trivial point, but it becomes crucial when the software needs to be displayed on a main screen at the front of a classroom using older projectors with poor resolution or color definition.

Concordancing With the AntConc Concordancer Tool

In the Osaka course, although the students are nonnative speakers, they use the English language at a fairly high level and thus can usually form complex grammatical sentences correctly. However, they often struggle when it comes to choosing appropriate vocabulary. For example, in the sentence below a student has used the phrase *a lot of*, which is inappropriate for technical writing in the target discipline of biochemistry.

> *There are* ***a lot of*** *antioxidative enzymes and low molecular compounds, which eliminate generated ROS.*

Many students find it difficult to realize the inappropriateness of *a lot of* as they are not familiar with the norms of the target discipline. In such cases, the Key Word In Context (KWIC) Concordancer Tool of AntConc can be effective. This tool enables the student to search for any word or phrase in a corpus, and then displays each hit on a separate line together with the words that appear next to it on the left and right (i.e., the surrounding context). AntConc also goes one step further and allows searches to contain a host of wildcards or even to form a regular expression (see http://rdues.uce.ac.uk/expressions.html/). Note that the KWIC lines usually have a fixed width in order to align the hits correctly. Although this is a standard design feature, one drawback is that words at the far left and right of the line will be terminated abruptly.

Returning to the above example, students at Osaka are asked to investigate the appropriateness of phrases, such as *a lot of*, by searching for them and similar phrases in their mini-corpus and identifying which expressions are used more frequently. The results of such a search are shown in Fig. 2 when carried out on a small corpus of 54 bioscience research abstracts. In this case, as expected, there are no occurrences of *a lot of* but eight occurrences of the

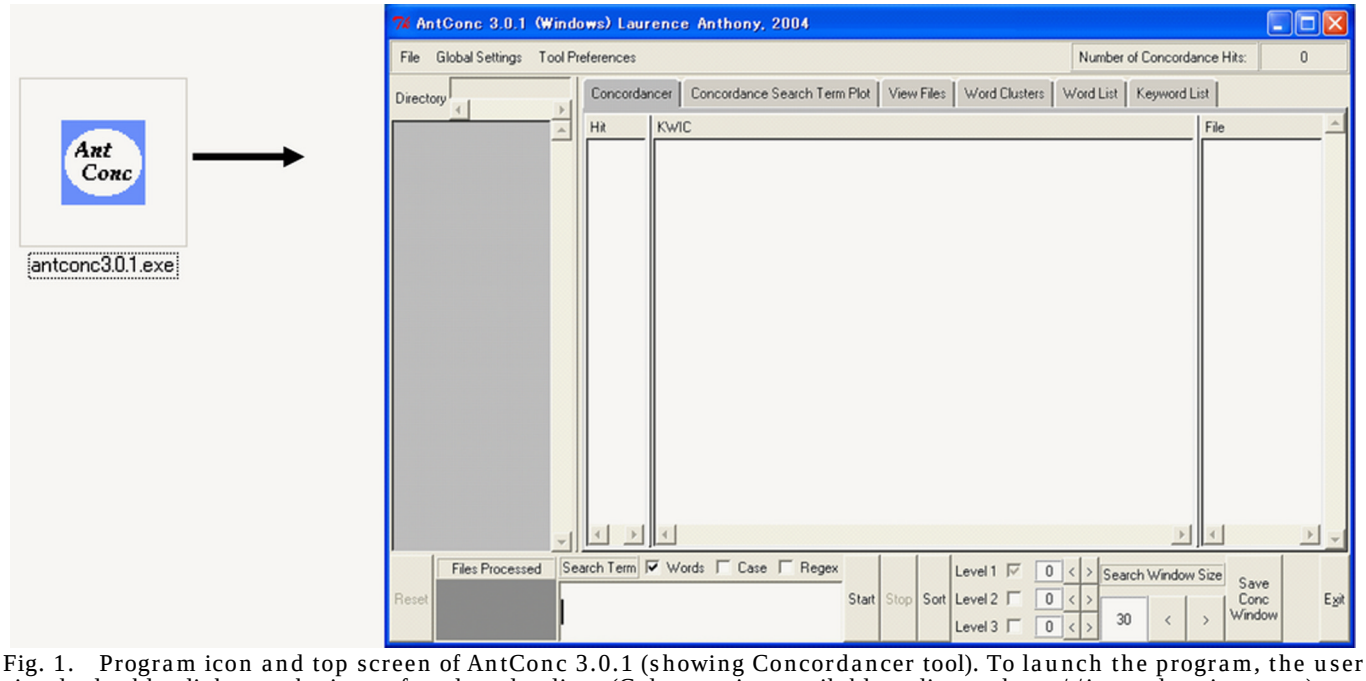


Fig. 1. Program icon and top screen of AntConc 3.0.1 (showing Concordancer tool). To launch the program, the user simply double clicks on the icon after downloading. (Color version available online at http://ieeexplore.ieee.org.)

more appropriate *many*, three of which correspond closely with the usage of *a lot of* in the example:

> *...oubtedly contribute to reveal many novel eukaryotic lineages, bu...* 13.txt
>
> *...rs controlling proteinases in many biological pathways. There is...* 28.txt
>
> *...tion and DNA damage response. Many Pin1 substrates are antigens...* 38.txt

When introducing the topic of word appropriateness in the classroom, a teacher may follow the set of procedures in Table II.

Concordancer tools like that in AntConc are not only useful in finding appropriate expressions, but as Sun and Wang describe, they also have been shown to facilitate the learning of vocabulary, collocations, grammar, and writing styles [13]. For example, research has shown new vocabulary can only be acquired through meeting words in diverse natural contexts and in varied situations [14], [15]. In a classroom with students from such diverse backgrounds, it is almost impossible for a teacher to find and explain a sufficient number of examples of a word or phrase to satisfy these conditions. On the other hand, with a specialized mini-corpus and a concordancer tool like the one in AntConc, a student can find and study a large number of examples in varied contexts and situations quickly and efficiently.

Finding Positional Information With the AntConc Concordance Search Term Plot Tool

One problem that students in the Osaka course have is not knowing **where** a term should be used in a particular text or section of text. For example, there is often confusion about if and where personal pronouns can be used in a research article. To resolve such issues, the Concordance Search Term Plot Tool of AntConc was added at the direct request of the Osaka University staff after they had used the software for six months.

The Concordance Search Term Plot Tool allows users to see where a search term appears by displaying each corpus file as a box in which the relative positions of search-term hits are shown as vertical lines. The importance of such a tool is demonstrated in Fig. 3. Here, a user has searched for the term *we* in the corpus of 54 bioscience research abstracts to find if and where it is used. The results clearly show that *we* is commonly used at the beginnings and ends of abstracts, where the authors discuss their previous research and their conclusions, respectively.

In addition to this usage, the tool has also been used at Osaka to determine, for example, if and where articles are used in research article titles, and if and where specialized terminology appears in full research articles.

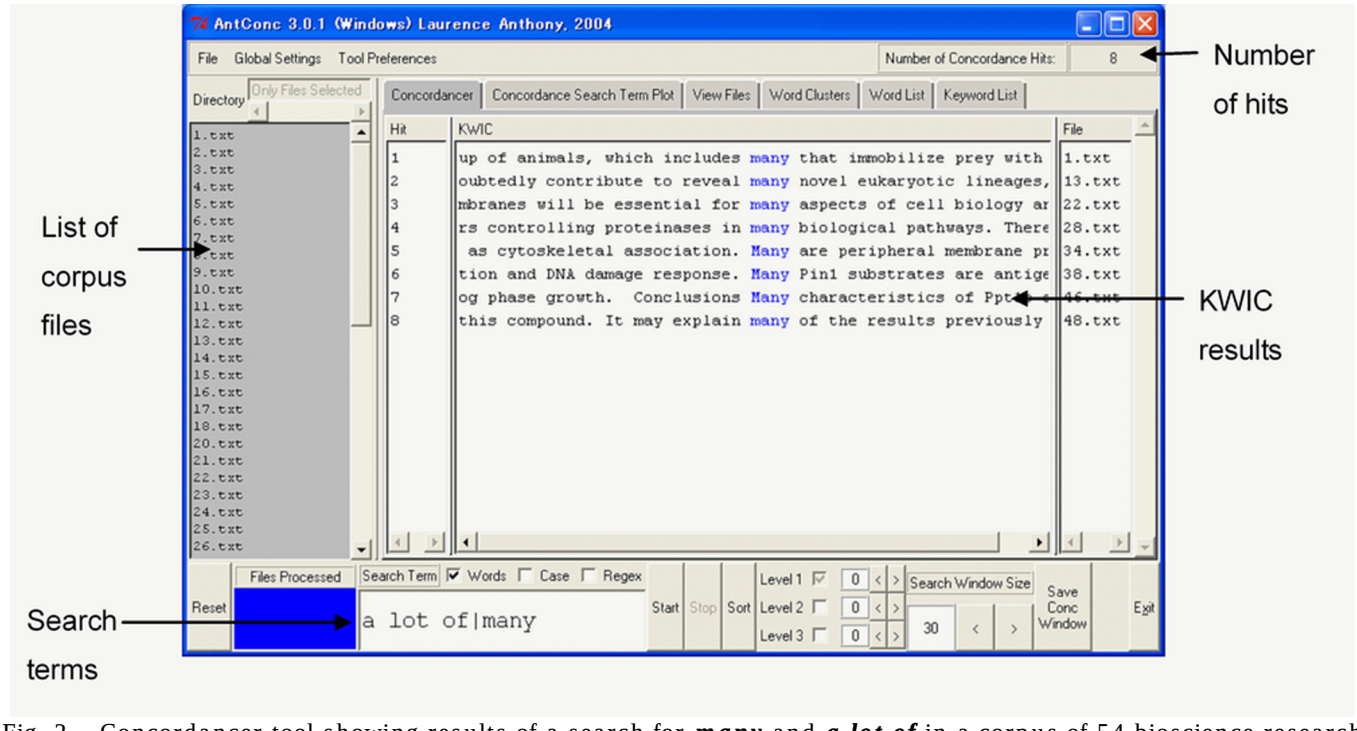


Fig. 2. Concordancer tool showing results of a search for ***many*** and ***a lot of*** in a corpus of 54 bioscience research abstracts. No occurrences of ***a lot of*** are found, but eight occurrences of ***many*** can be seen. (Color version available online at http://ieeexplore.ieee.org.)

TABLE II
POSSIBLE PROCEDURES TO FOLLOW IN A TECHNICAL WRITING CLASS WHEN INTRODUCING WORD APPROPRIATENESS

Timing	Procedures
Preparation	Ask students to prepare a mini-corpus of around 50~100 target texts in their discipline through web-searches or online databases of well-established, high quality work.
	Ensure all students have access to *AntConc* on their computer desktops and can navigate to their prepared corpus.
Pre-class Exercise	Have students prepare a short sample of their writing and have them note down words and phrases they feel unconfident about.
	Teacher-edit or peer-edit a sample of the students' writing, highlighting (but not correcting) problems of word appropriateness.
In-class Exercise	Give a general introduction on word appropriateness and the importance of field-specific differences in word choice. Also, offer guidelines on how to use a thesaurus or dictionary to find alternative words and phrases.
	Have students open their corpora in *AntConc* and search for the words and phrases highlighted or thought to be inappropriate in their sample texts using the Concordancer Tool.
	Ask students to consider their results in terms of frequency of occurrence, distribution across corpus texts (via the Concordance Plot Tool), and size of corpus analyzed. For example, if a term appears commonly in many texts suggest that they can feel confident that they are using the term appropriately. On the other hand, if the term is used infrequently or in only a small number of texts, ask them to consider the possible inappropriateness of the term in their field.
	If a search term is deemed potentially inappropriate, ask students to search for alternative forms in their corpora based on their thesaurus or dictionary findings.
	Ensure that students undertake the exercise as investigators who are attempting to "discover" features of language in their target disciplines either individually or in a group. In this context, the teacher should not provide "answers", but instead provide suggestions and feedback on the results generated.

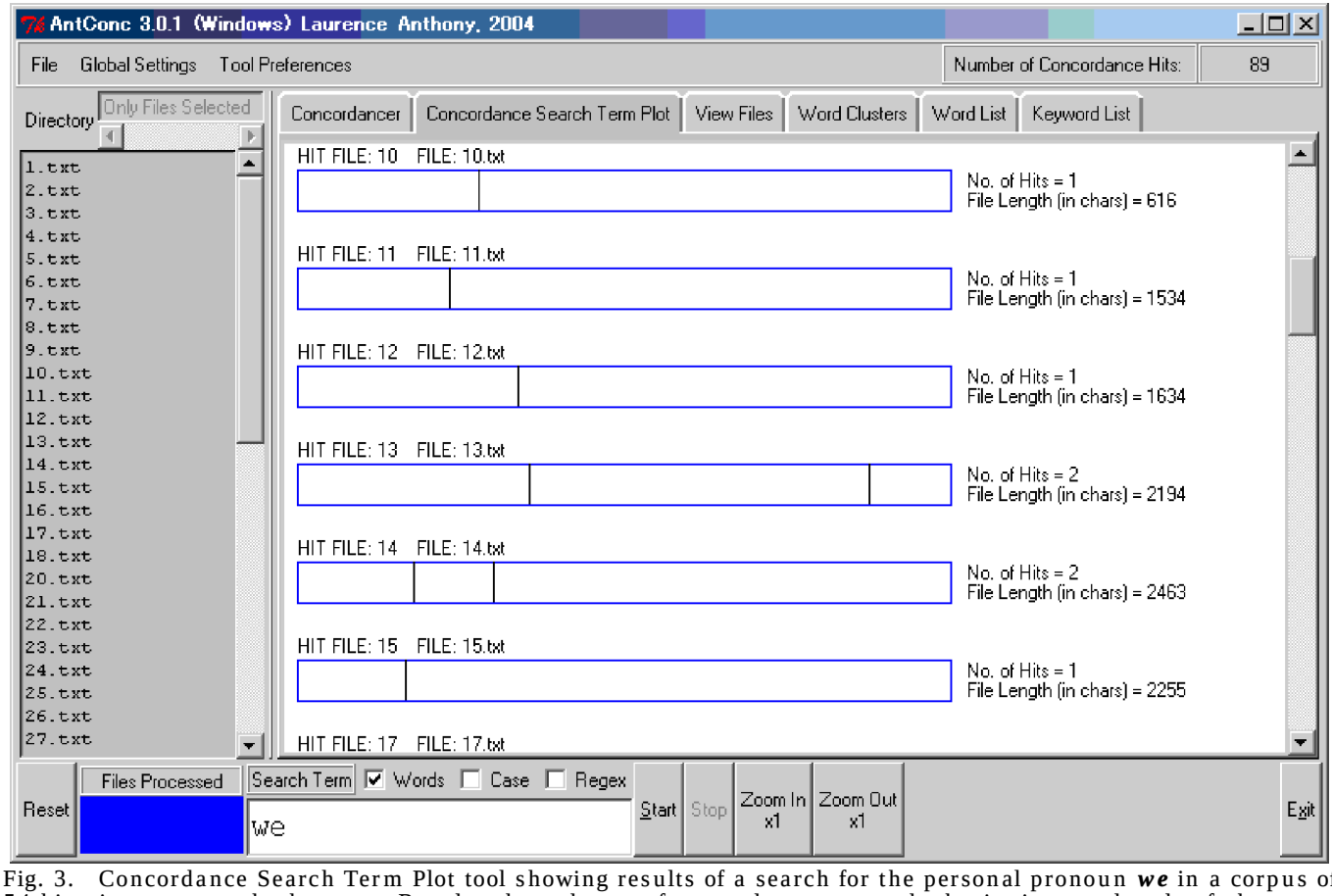


Fig. 3. Concordance Search Term Plot tool showing results of a search for the personal pronoun ***we*** in a corpus of 54 bioscience research abstracts. Results show that ***we*** frequently occurs at the beginnings and ends of abstracts. (Color version available online at http://ieeexplore.ieee.org.)

VIEWING COMPLETE TEXTS WITH THE AntConc VIEW FILES TOOL

When conducting corpus investigations, it is always useful for the learner to be able to view the complete texts. This is because there are many occasions when the user will need to view search results within a larger context than that provided by a KWIC concordance line. To enable users to view texts and also search for any substring, word, phrase, or regular expression in a single corpus file, the View Files Tool was added to AntConc.

Viewing single texts after generating a set of concordance lines is a common action, so the Concordancer Tool and View Files Tool were developed for smooth interchangeability through the use of hyperlinks. In this way, whenever a search term in a KWIC concordance line is clicked, the View Files Tool is automatically activated and programmed to display the hit in its original file. Similarly, if a hit in the View Files Tool is clicked, all occurrences, in all files of that hit, are displayed via the Concordance Tool.

An example of the View Files Tool in action is shown in Fig. 4, where the results of clicking on one of the hits for *many*, as described in the section titled Concordancing with the AntConc Concordancer Tool, are displayed.

GENERATING WORD/KEYWORD LISTS WITH THE AntConc WORD LIST TOOL AND KEYWORD LIST TOOL

One of the first things that students are asked to do in the Osaka course after creating a new corpus is to generate a list of all the words in the corpus together with their frequencies using the Word List Tool of AntConc. Word lists are useful as they immediately suggest interesting areas for investigation and highlight problem areas in a corpus. For example, by listing all the words used in a bioscience corpus, students can quickly appreciate the importance and role of content words, such as *expression*, *purification*, and *characterization*. Bowker and Pearson describe how word lists can also be used to find families of related word forms and lemmas in a corpus [16].

Hockey states that an ideal word-list generation program should be able to sort words into alphabetical or frequency order [17]. The Word List Tool of AntConc offers these features and two added features: reverse ordering and counting words based on their stem forms. However, when generating a word list, students are usually disappointed to find that almost all the high-frequency words are function words. To ignore these in the list, a so-called stop list can be defined in the Word List Tool preferences, or alternatively the user can specify the reverse of a stop list (i.e., a list of only the words that should be counted). However, even a raw word list can be revealing, for example,

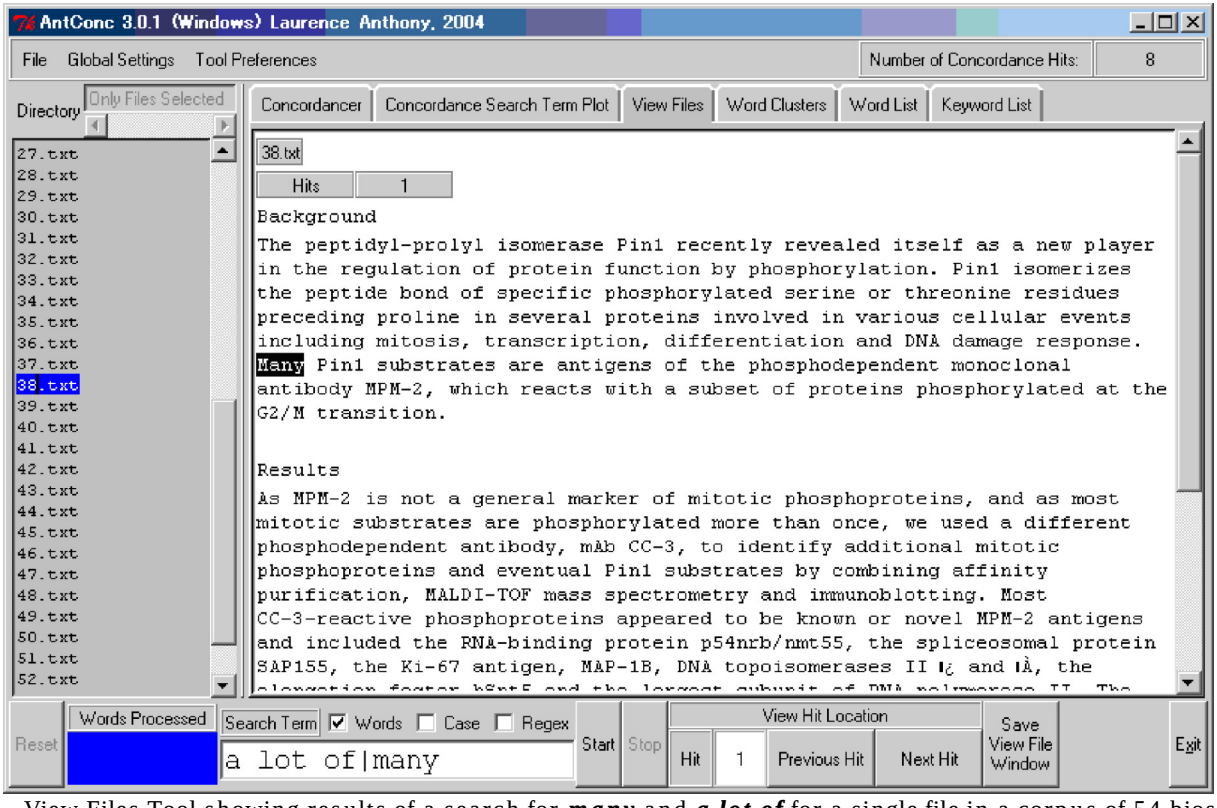
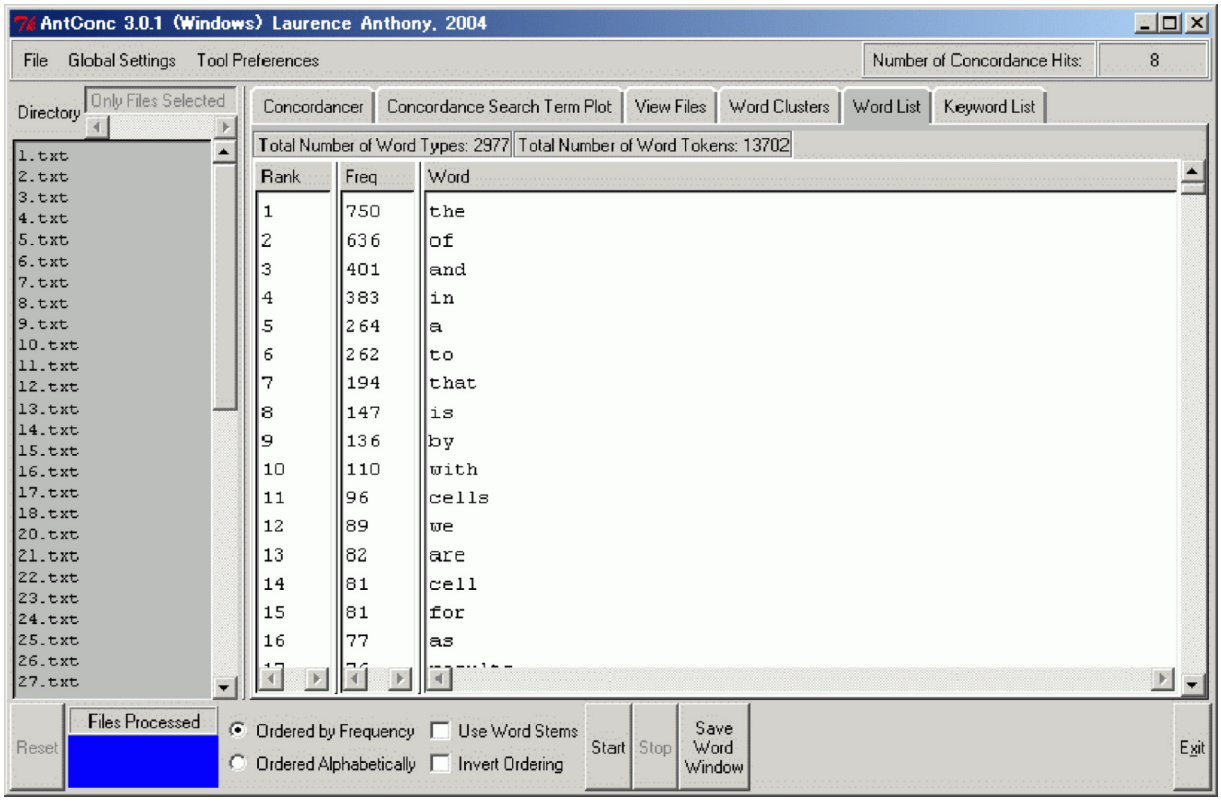
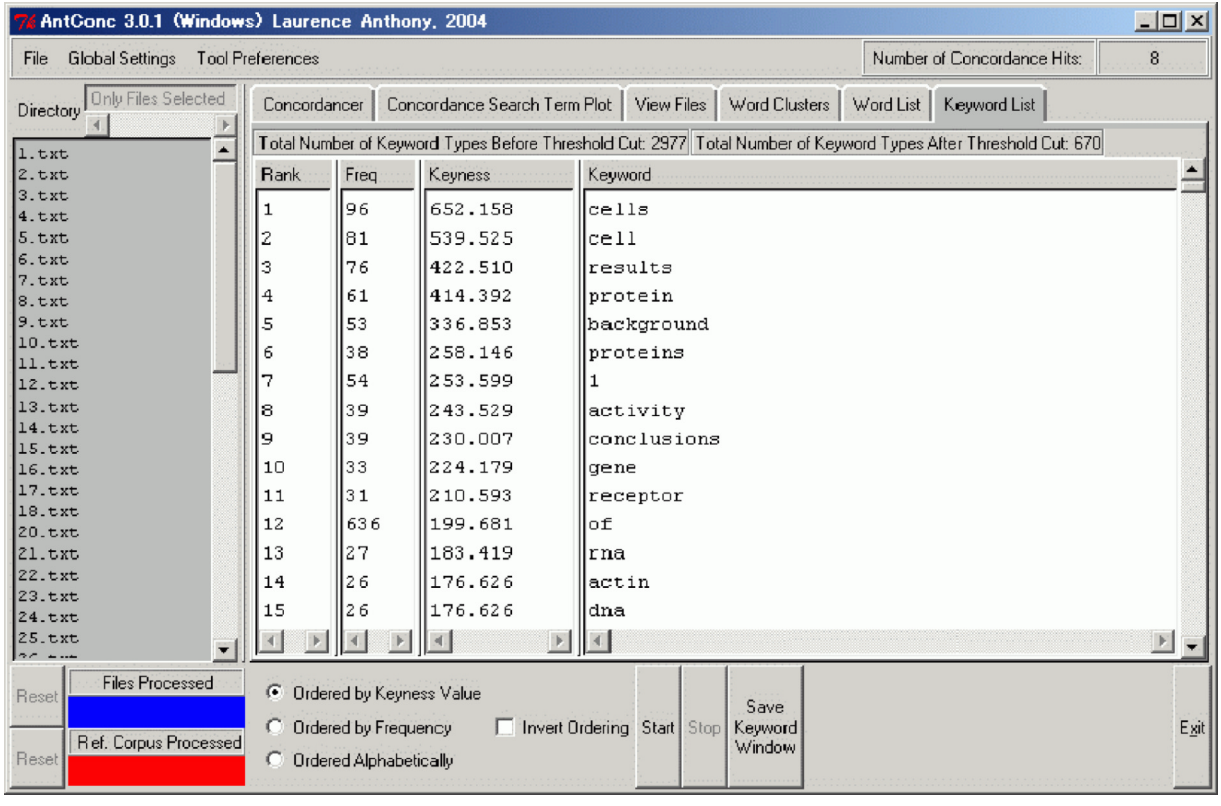


Fig. 4. View Files Tool showing results of a search for ***many*** and ***a lot of*** for a single file in a corpus of 54 bioscience research abstracts. No occurrence of ***a lot of*** are found, but the one occurrences of ***many*** is highlighted. (Color version available online at http://ieeexplore.ieee.org.)



(a)



(b)

Fig. 5. (a) Word List Tool showing the words in a corpus of 54 bioscience research abstracts ranked by frequency. The word ***the*** at the top of the list is shown to have the highest frequency occuring 750 times in the corpus. (b) Keyword List Tool showing the keywords in a corpus of 54 bioscience research abstracts ranked by the log-likelihood measure. The word ***cells*** at the top of the list is shown to be the most unusually frequent word in the target corpus compared with its occurrence in a corpus of four English novels. (Color version available online at http://ieeexplore.ieee.org.)

in the case of partial texts, such as research article abstracts or titles. As Fig. 5(a) shows, in the small corpus of 54 bioscience research abstracts, students at Osaka are often surprised to find that one of the most common words is *of*. This reflects a tendency in this field and many others to use long noun phrases, as the following example illustrates:

> *Investigation **of** the structure **of** the ribosomal subunits in complex with different antibiotics can reveal the mode **of** inhibition **of** ribosomal protein synthesis.*

Experienced users of corpus tools will know that word lists usually tell us little about how important a word is in a corpus. Therefore, AntConc also offers a Keyword List Tool, which finds words that appear unusually frequently in a corpus and compares them with the same words in a reference corpus specified by the user. The Keyword List Tool calculates the "keyness" of words using either the chi-squared or log likelihood statistical measures and offers the user the option of displaying or hiding unusually infrequent keywords (or negative keywords) in the preferences window [18]. In Fig. 5(b), a keyword list for the corpus used in Fig. 5(a) has been generated using a reference corpus comprised of four English novels. In class, where students will have their individual corpus of texts, such a list can form the basis of a personalized list of essential vocabulary that should be learned.

Investigating Multiword Patterns With the AntConc Word Clusters/Bundles Tool

Research has shown that collocations and other multiword units, such as phrasal verbs and idioms, are particularly difficult for learners to acquire [19]. Their importance is even greater if the learner is working with texts in a highly technical or scientific field, as the lexical unit is usually longer than a single word [16]. Students in the Osaka course, for example, are often unsure whether to combine terms such as *gene* and *transfer* in the form *transfer of genes* or the simpler pattern *gene transfer*. Surprisingly, these multiword units have received little attention in most CALL programs, perhaps due to the difficulty in identifying and ordering them in a systematic way for the learner [19].

In AntConc, collocations and multiword units can be investigated using the Word Clusters Tool. This tool displays clusters of words centered on a search term and orders them alphabetically or by frequency. The search terms can be specified as a substring, word,

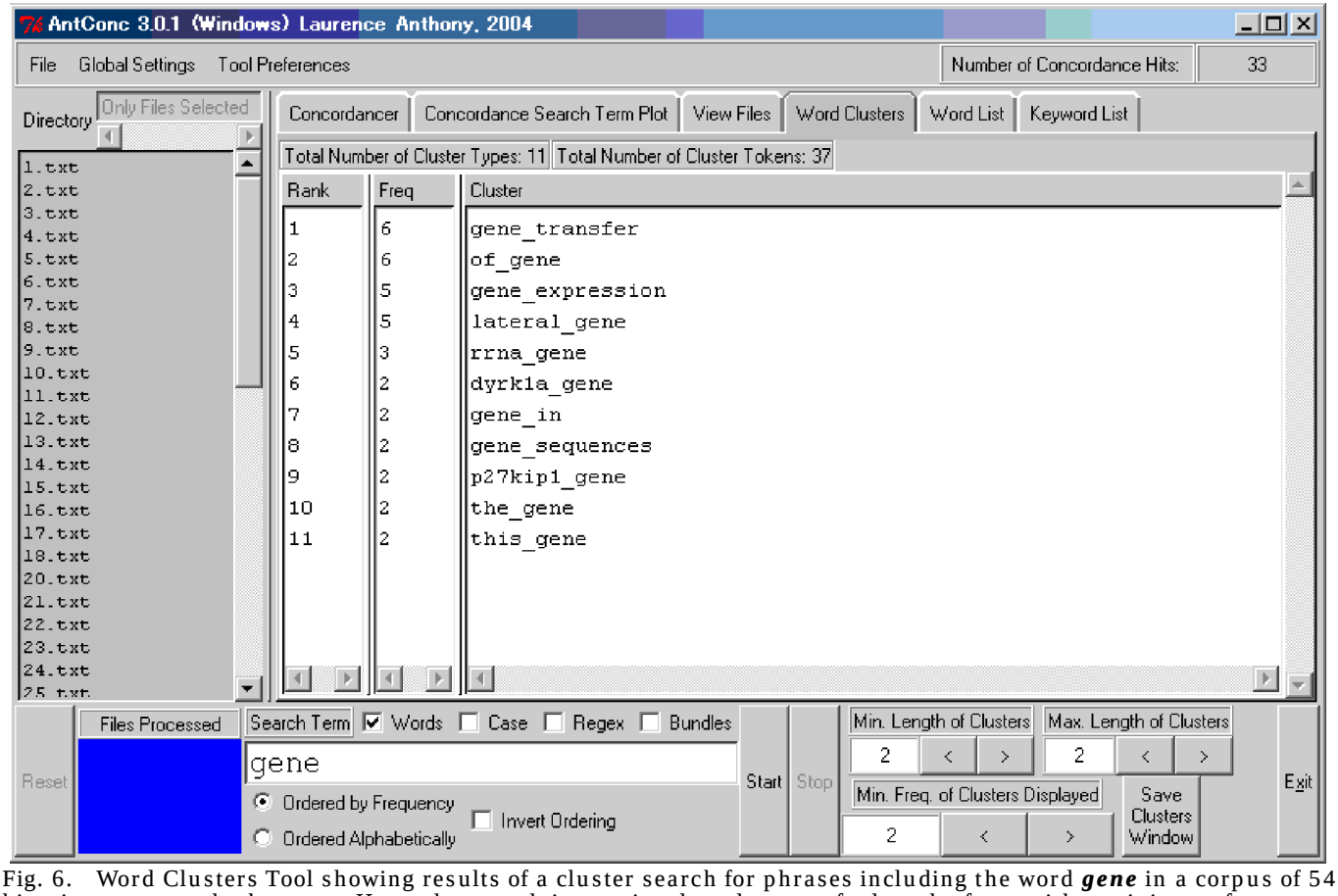


Fig. 6. Word Clusters Tool showing results of a cluster search for phrases including the word ***gene*** in a corpus of 54 bioscience research abstracts. Here, the search is restricted to clusters of a length of two with a minimum frequency of two. (Color version available online at http://ieeexplore.ieee.org.)

phrase, or regular expression as in the Concordancer, Plot, and View File tools. The number of additional words to the left and right of the search term can also be specified, and it is possible to set a minimum frequency threshold for the clusters generated.

As shown in Fig. 6, using the Word Clusters Tool reveals that in the corpus of 54 bioscience research abstracts, the word *gene* is commonly combined with the word *transfer* in the form *gene transfer* and is grouped with other words in a similar way, as in *gene expression* and *gene sequences*. By combining this tool with the other tools in AntConc, students can begin investigating why and how these phrases are used in research articles and can learn, for example, if they can be considered as countable or uncountable.

An alternative way to search for multiword sequences is to find lexical bundles, which are equivalent to n-grams, where n usually varies between two and five words [20]. Few corpus analysis programs offer this feature, but AntConc includes lexical bundle searches as an option in the Word Clusters Tool [1]. Calculating all the lexical bundles for a particular set of criteria can take a great deal of time. Therefore, as in all other tools in the program, if a student finds the processing is taking too long, it can be halted by simply clicking on the stop button.

Limitations and Future Developments of AntConc

Concordancers can be divided into two main types: (1) those that first build an index that is used for subsequent search operations, and (2) those that act directly on the raw text [17]. The first of these has the advantage of operating on large corpora and producing results more quickly after the index has been built. On the other hand, it tends to be less flexible than the second type, especially if the user often switches or modifies the target corpus for a particular need.

AntConc fits into the second category, performing all processing on the raw data files and storing results in active memory. For this reason, its use is limited to small specialized corpora and suffers slightly in terms of speed. Nevertheless, as McEnery and Wilson note, a trend in corpus linguistics over the past few years is the increased interest in very small, highly specialized corpora [7]. Small corpora can be used for many different purposes, as exemplified by Ghadessy et al., and are particularly effective when teaching technical writing, as in the Osaka course [21]. Also, as mentioned earlier, because AntConc performs all operations in active memory, it does not need to be installed on the target computer and thus overcomes many system management issues.

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. To date, this feature has not been added to AntConc as students have tended to find such tables difficult to interpret. On the other hand, an increasing number of users have requested this feature, so I will be including it in the next major upgrade of the program.

Some programs also 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 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 [17]. In AntConc, the results in all display windows can be easily copied and pasted directly into a spreadsheet program using simple keyboard shortcuts. However, in the program's next release, I am considering whether to add some simple statistics as a direct result of feedback from a user in the US.

As one of AntConc's users in Japan has pointed out, one of the weakest areas of the program is its handling of annotated data, such as part-of-speech information. Although AntConc offers a simple way to view or hide embedded tags, much more sophisticated methods need to be implemented if the full power of annotated data is to be realized. Implementing such methods is an active area of current research.

In addition to the features mentioned above, in the next release of AntConc, users will be able to sort word lists alphabetically from both the beginning and end of words, a feature recommended by Hockey [17]. Also, I plan to create a detailed user manual and accompanying tutorial video, where each tool's operation will be explained with concrete examples and a step-by-step guide. Although these are not strictly necessary (as the program operation is largely intuitive), for users with little experience of corpus linguistics or computers, they could be valuable aids in the early weeks of a course, such as that at Osaka University.

Discussion and Conclusions

AntConc is a lightweight, fully-featured and easy-to-use corpus analysis toolkit that has been shown to be effective in the technical writing classroom [11]. Although it does not include every tool and feature included in popular commercial applications, it offers many of the essentials needed for corpora analysis, with the added benefits of an intuitive interface and a freeware license.

Feedback from users around the world has clearly shown that AntConc can also be used successfully as a tool with both native and nonnative speakers. Two recent examples of comments I have received are given below.

User in Japan:

I just want to thank you for AntConc. I've used it a lot personally and just recently with a class using the COBUILD dictionary w/CD-ROM as a textbook. With the corpus on there and AntConc they could investigate how words were used and identify collocates from the Wordbank on the CD that wouldn't have been possible otherwise. Thanks a lot.

User in Australia:

Just dropping you a note of thanks for the impressive AntConc. I came across AntConc while looking for tools to analyze my writing efforts—a salutary exercise.

AntConc is not limited only to classroom applications. On the contrary, the software has potential uses in many areas of professional communication. As Hunston describes, tools such as AntConc can help writers of dictionaries, coursebooks, and grammar books by revealing information about the relative frequencies between words, meanings, and usages [3]. They can also help in selecting suitable examples to illustrate typical word and phrase usage. Kenny gives examples showing the importance of corpus linguistics tools for translators of German texts, and Anthony demonstrates how tools can be used to help university staff in writing recommendation letters for prospective graduate school candidates [22], [23]. In short, for any context in which a corpus of electronic data can be collected, the potential uses of software such as AntConc are nearly endless. For example, it is a simple matter to envision using corpus linguistics tools in the writing of grant proposals, automobile recall notices, product user manuals, and web-based help pages, as well as in spoken contexts, such as in the analysis of the language used in presentations, dialogues between pilots and air traffic controllers, and even doctor-patient interactions.

The rapid growth in corpus linguistics will continue to have an impact on many fields related to language and professional communication in coming years. As discussed in this paper, I hope that by continuing to develop and refine AntConc, this software can also continue to serve the varied needs of learners, teachers, researchers, and practitioners interested in corpus linguistics techniques.

ACKNOWLEDGMENT

This work was supported by a Grant-in-aid for Scientific Research by the Japan Society for the Promotion of Education, Science, Sports and Culture, Japan (16700573), and by a Waseda University Grant for Special Research Projects, Japan (2004B-861).

REFERENCES

[1] D. Coniam, "Concordancing oneself: Constructing individual textual profiles," *Int. J. Corpus Linguistics*, vol. 9, no. 2, pp. 271–298, 2004.
[2] C. A. Chapelle, *Computer Applications in Second Language Acquisition: Foundations for Teaching, Testing, and Research*. Cambridge, UK: Cambridge Univ. Press, 2001.
[3] S. Hunston, *Corpora in Applied Linguistics*. Cambridge, UK: Cambridge Univ. Press, 2002.
[4] T. Johns, "Contexts: The background, development and trialling of a concordance based CALL program," in *Teaching and Language Corpora*, A. Wichmann, S. Fligelstone, T. McEnery, and G. Knowles, Eds. London, UK: Longman, 1997, pp. 100–115.
[5] J. M. Swales, *Research Genres*. Cambridge, UK: Cambridge Univ. Press, 2004.
[6] J. M. Levis and G. M. Levis, "A project-based approach to teaching research writing to nonnative writers," *IEEE Trans. Prof. Commun.*, vol. 46, no. 3, pp. 210–220, Sep. 2003.
[7] T. McEnery and A. Wilson, *Corpus Linguistics. An Introduction*, 2nd ed. Edinburgh, UK: Edinburgh Univ. Press, 2001.
[8] D. Biber, S. Conrad, and R. Reppen, *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge, UK: Cambridge Univ. Press, 1998.
[9] G. Kennedy, *An Introduction to Corpus Linguistics*. London, UK: Longman Press, 1998.
[10] C. F. Meyer, *English Corpus Linguistics: An Introduction*. Cambridge, UK: Cambridge Univ. Press, 1998.
[11] J. Noguchi, "A genre analysis and mini-corpora approach to support professional writing by nonnative English speakers," *English Corpus Studies*, vol. 11, pp. 101–110, 2004.
[12] C. Lonfils and J. Vanparys, "How to design user-friendly CALL interfaces," *Comput. Assisted Lang. Learning*, vol. 14, no. 5, pp. 405–417, 2001.
[13] Y. C. Sun and L. Y. Wang, "Concordancers in the EFL classroom: Cognitive approaches and collocation difficulty," *Comput. Assisted Lang. Learning*, vol. 16, no. 1, pp. 83–94, 2003.
[14] T. Cobb, "Breadth and depth of lexical acquisition with hands-on concordancing," *Comput. Assisted Lang. Learning*, vol. 12, no. 4, pp. 345–360, 1999.
[15] K. E. Nitsch, "Structuring Decontextualized Forms of Knowledge," Ph.D. dissertation, Vanderbilt Univ., Nashville, TN, 1978.

[16] L. Bowker and J. Pearson, *Working With Specialized Language: A Practical Guide to Using Corpora*. London, UK: Routledge, 2002.
[17] S. Hockey, "Concordance programs for corpus linguistics," in *Corpus Linguistics in North America: Selections from the 1999 Symposium*, R. C. Simpson and J. M. Swales, Eds. Ann Arbor, MI: Univ. Michigan Press, 2001, pp. 76–97.
[18] A. Kilgarriff, "Comparing corpora," *Int. J. Corpus Linguistics*, vol. 6, no. 1, pp. 97–133, 2001.
[19] N. Nesselhauf and C. Tschichold, "Collocations in CALL: An investigation of vocabulary-building software for EFL," *Comput. Assisted Lang. Learning*, vol. 15, no. 3, pp. 251–279, 2002.
[20] D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan, *Longman Grammar of Spoken and Written English*. London, UK: Longman, 1999.
[21] M. Ghadessy, A. Henry, and R. L. Roseberry, *Small Corpus Studies and ELT: Theory and Practice*. Amsterdam, The Netherlands: John Benjamins, 1996.
[22] D. Kenny, "Translators at play: Exploitations of collocational norms in German, English translation," in *Working with German Corpora*, D. Dodd, Ed. Birmingham, UK: Univ. Birmingham Press, 2000, pp. 143–160.
[23] L. Anthony, "Concordancing with AntConc: An introduction to tools and techniques in corpus linguistics," presented at the Corpus Linguistics Workshop 2005 (ICCL 2005), Tokyo, Japan, Nov. 25–27, 2005.

Laurence Anthony received the M.A. degree in TESL/TEFL and the Ph.D. in Applied Linguistics from the University of Birmingham, UK, and the B.Sc. degree in mathematical physics from the University of Manchester Institute of Science and Technology (UMIST), UK. He is Associate Professor in the School of Science and Engineering at Waseda University, Japan, where he teaches technical reading, writing, and presentation skills, and is coordinator of the technical English program. His primary research interests are corpus linguistics, computer assisted language learning, educational technology, genre analysis, and natural language processing.