Complementing University EFL Science Courses: A CALL Multimedia Concordance Design for Vocabulary Comprehension

Richard Gilbert
Ryoji Matsuno

Abstract

Although computer-based concordance programs are available for the researcher, a multimedia concordance program has not yet been authored that is specifically designed to both meet the needs of academic, learner-centered reading programs, and which seeks to increase student comprehension of scientific texts in ESP (English for specific purposes) settings. This paper presents a computer assisted language-learning (CALL) software design and rationale for a tripartite concordance and lexis that reflects the three major groupings of English vocabulary deemed particularly relevant for English for academic purposes (EAP) science study, as well as other academic fields. The ultimate goal of the program under development is to enable vocabulary study to be further integrated into EAP reading classes, with the potential of accelerating students' comprehension of the vocabulary, texts, and contexts in which the vocabulary items are contained. The CALL module discussed in this paper is the central module of a projected multi-module, multimedia CALL reading program, designed to provide (a)vocabulary, (b)reading materials, (c)student assessment, (d)test creation, (e)data collection/compilation of

study and test histories, and (f)grading functions for the teacher (or independent student). Teachers and departments will be able to utilize the total program as an educational tool in designing distance learning and student self-study courses, as well as integrate the program into courses which make use of computer labs. The concordance module discussed in the following paper can be utilized as a 'standalone' program.

Introduction

Paradigm Shifts

In a recent comprehensive overview of the English for specific purposes (ESP) field, CALL programs, on the whole, were reviewed in a dim light: "CALL programs are largely linear, constructed along certain thought patterns, with a single or limited response" (Dudley-Evans and St. John, 1998, p.184). A partial explanation may be offered by considering three main constraints confronting CALL program design: (a)technology paradigm shifts; (b)non-standardization/evolution: (i)of operating systems (OSs), (ii)media communication standards (codecs), (iii)development environment problems; and (c)pedagogy/content-design paradigm shifts. A brief explication follows. Constraints (a)and (b): Computer processing power is doubling every 18 months-with the result that hardware is generally obsolete within three years. It is only recently (in the last three-year computer "era"), that multimedia CD-ROM computers have become affordable devices (under $1,000 USD) able to interact with users utilizing several simultaneous media, in real-time. Thus, the lag between hardware evolution and software development, combined with the rapidity of evo-

lution of OSs, hardware, and processing power has created a nightmare for content developers. OS platforms and standards are as the shifting sands of the desert; development of educational software able to take advantage of the latest multimedia technologies has therefore been slowed. When programs do appear they may disappear again quickly: "Because multimedia technologies are evolving rapidly, new software often becomes obsolete within half a year" [R. Matsuno, Trans.] (Kikue and Iiyoshi, 1996, pp.109 - 110).

We have all become familiar with, and probably use daily, powerful software written and yearly revised by large teams of highly - paid professionals, working for companies with profits in the billions of dollars. By contrast, educational software is often produced non - commercially by one or a small group of individuals with far more limited resources. Until recently, object - oriented (OOP) software development environments catering to individual developers were not powerful enough to create interactive multimedia applications unless accompanied by labyrinthine hand - coding tasks and debugging. To the present, any designer of interactive educational multimedia programs is confronted with an undertaking that is more likely to be measured in years than months before the software program will become mature - especially when the necessity for revision, field - testing, "tweaking," and upgrading is considered. For these reasons, viable CALL programs have tended towards the 'minimalistic' end of the software spectrum.

Constraint (c): Another reason for a "single or limited response," as well as a narrowness of focus in CALL programs, may have to do with an historical paradigm of specialization, in which academics traditionally focus their research interest within a particular field or sub - field. Increasingly, computer program design, especially interactive multimedia design, demands a generalist paradigm involving multiple

disciplines, and a collaborative/team approach (cf. Alexander, 1999). This new paradigm presents challenges within an academic environment, and program design suffers when these challenges are not well met. Multiple dimensions of learner control and program options related to learner autonomy are design concepts which should be considered as principles basic to CALL design (Berge & Collins, 1995a; Soo, 1998; Steinberg, 1989), but which are alien to the traditional classroom. Planning for learner autonomy within a CALL program and integrating CALL within the classroom demand novel pedagogical approaches. McClintock (1992) and others (e.g. the Institute for Learning Technologies, Columbia University), have written extensively on this subject. Novel methodological approaches relating to the integration of computers in classrooms as educational tools challenge current pedagogical and content-design paradigms:

> Computer-mediated communication (CMC) promotes a type of interaction that is often lacking in the traditional teacher-based classroom. It allows learners the freedom to explore alternative pathways - to find and develop their own style of learning . . . . Computer technologies allow professionals to share with students tools that we use daily. Further, as educators, we can provide guidance to help students develop meaningful ways to construct their own knowledge, much as we ourselves do. Technology enables us to implement these new visions . . . [while not excluding] experts in [various] fields of inquiry from the collaboration (Berge and Collins, 1997b).

Currently, we are experiencing yet another technological paradigm shift, as computers are not only becoming more portable but are now

becoming wearable (i.e. "smart clothing"). The conception that a computer is something that either sits on a desktop or is carried about like a fragile jewel in a case, is changing. The next generation of hardware promises to create more options and opportunities for human - computer interface (HCI) design, as the internet, multimedia, software, and telecommunications become further integrated into one wireless, portable device. Forearm - mounted and heads - up display devices are beginning to appear, along with unfoldable, super - thin keyboards, monitor screens, and up to 24 - hour battery - powered operating times. Later this year, the fourth (M.I.T. - sponsored) international symposium on wearable computing (IEEE ISWC 2000: http://www.media.mit.edu/Wearables/) will be held. Along with smart clothing, the 'e - book computer' is becoming a reality. This, from a recent article: "coming soon . . . the 'Tablet PC,' a computer that looks like a flat - screened magazine for reading e - books and which lets you annotate the 'pages' in your own handwriting and even lets you search through your handwritten notes" (Mieszkowski, 2000). It is expected that in the next few years, novel types of portable computers will become more plentiful, available to students at a reasonable cost, and useful within L2 language - study environments. Though the timing of the acceptance and spread of such innovations must be imprecise, it is reasonable to assume that three of the most severe disadvantages to multimedia CALL education - (a)cost, (b)lack of portability, and (c)the necessity for complex institutional infrastructures, may be largely overcome within this decade. Although the CALL design under discussion currently demands a computer lab environment in an academic classroom setting, strategic plans foresee a more ubiquitous EAP online classroom in the future.

Accessing New Vocabulary: How Much Of A Good Thing?

Over the past 15 or so years, there has been increasing interest in the role that vocabulary comprehension plays in improving reading ability in English for academic purposes (EAP). Prior to the mid-80s, few studies had yet been undertaken to analyze the vocabulary needs of students studying in technical fields. Three factors brought about both a renewed interest in vocabulary studies and the ability to formulate novel research approaches within the context of EAP. First, through the growing international relevance and influence of the English for Specific Purposes (ESP) movement, an especial focus was placed on ESL/EFL learners who needed to improve ability within the context of a specific endeavor, field, or vocation of study (Hutchinson & Waters, 1987). The second factor was technological, involving the growth of the personal computer (PC) market: the availability of inexpensive, powerful PCs for the researcher, teacher, and student. A third factor was the development of research methodologies utilizing corpus, collocation, and concordance software, and the evolution of the ease of use, affordability, and power of these tools.

Current vocabulary research has enabled researchers to determine, with a specificity heretofore logistically impossible, the types of vocabulary, and 'percents' of known vocabulary, necessary for reading comprehension within particular areas of study. In designing a CALL program for vocabulary comprehension as an adjunct to academic readings in science classes, there exists a rationale for (a) dividing vocabulary into word families, (a word family is signified by a "header" word, followed by all inflected and derived forms of that headword. For example, the headword analyze contains the word family: analyzable, analysis, analyzed, analyzing, unanalyzable, re-analyzing, non-analyzable, etc.); and, (b) applying the results of corpora re-

search by grouping word families into three distinct levels, based upon frequency of usage and range: "The division of the vocabulary of academic texts into three levels of [1]general service or basic vocabulary, [2]sub-technical vocabulary, and [3]technical vocabulary is a commonly made distinction" (Nation, 1999, p.274). These three groups represent a hierarchy of decreasing frequency and range. A student's comprehension or lack of comprehension of word families in these respective levels provides a strong indication of English reading ability, especially as related to academic texts—and can be determined by such measures as the Vocabulary Levels Test.

> In a study of the Vocabulary Levels Test (Nation, 1983; 1990) found that second language learners' scores on the various levels of the test decreased from the high frequency levels to the lower frequency levels. That is, the test had strong implicational scaling. Laufer and Nation (1999) found a similar effect for productive knowledge (Hu and Nation, 1999).

It should be noted that vocabulary knowledge and reading comprehension are very closely related to each other (Nation, 1999; Stahl, 1990). In fact, the correlation is high and reliable enough that TOEFL researchers at one point considered whether both vocabulary items and reading items needed to be included in that test, "and if not, which of the two could be dispensed with" (Read, 1997, p.305). Three vocabulary lists are utilized in our concordance design. They are: 1)the 2000 highest frequency word families included in West's "General Service List of English Words (GSL)," (1953), further sub-divided into two subgroups of 1000 word families; 2)Coxhead's "Academic Word List (AWL)," (1998), a multi-disciplinary academic word list of 570 word families,

sub-divided into 10 sub-groups, by frequency ("academic words" are equivalent to the term "semi/sub-technical vocabulary" in ESP nomenclature); and 3)technical vocabulary, representing words or terms that are largely field-specific.

We might ask, why select out particular, limited vocabulary sets for study, while ignoring the bulk of the English vocabulary? The answer lies in a real-world equation having to do with the number of word families in English, the disparity of vocabulary knowledge between typical undergraduate L1 and L2 English users, and the limited time available to educate students in order to achieve academic goals in English, e.g., reading essays and research papers in their academic field, the overwhelming majority of which are in English:

> Even many journals of smaller nations' scientific societies, like those of Slovenia for example, publish also in English. When abstracted more widely these are then accessible to a world audience. For a scientist to publish in a language other than English therefore is increasingly to cut herself off from the world-wide community of scientists who publish in English. The work may then be ignored simply because it is published in a language unknown to the rest of the world (Wood, 1997).

Webster's Third New International Dictionary contains approximately 114,000 word families, excepting proper names (Goulden, Nation and Read, 1990); a number well beyond the knowledge of most L1 or L2 students. Though past measurements of vocabulary size have been found to be erroneous (Nation, 1999), recent reliable studies (Goulden, Nation and Read, 1990; Zechmeister, Chronis, Cull, D'Anna

and Healy, 1995) indicate that educated native English speakers know about 20,000 word families; the educated undergraduate L2 learner of English may have 10% - 15% of this vocabulary, a comprehension of 2,000 - 3,000 or so word families—placing them at a great disadvantage by comparison, especially in that "to read with a minimal disturbance from unknown vocabulary, [English] language users probably need a vocabulary of 15,000 to 20,000 words" (Nation, 1999, p.15).

Researchers have attempted to determine what percent of the vocabulary in a given text needs to be comprehended by an L2 learner of English in order to (a)adequately comprehend the text, (b)comprehend a new vocabulary item by context, and (c)read comfortably/read for pleasure. Research (Laufer 1992, Liu Na and Nation 1985, Hu and Nation 1999) suggests that at least 95% of the running words in a text need to be known to the learner, in order for unknown vocabulary to be comprehended by context. This 95% figure, which indicates that a reader will encounter, on average, one unknown word for every two lines of (10 words average per line) text, and 20 unknown words on a (400 - word average) page, seems as well to be a lowest threshold for successful reading comprehension. Laufer (1989), examining successful readers of an English for Academic Purposes text in the First Certificate in English exam, found that scores in the 95% percentile and above had a higher number of successful readers. Readers who scored at 90% (one unknown word per line of text) did not do significantly better or worse than others scoring up to 94%; these scoring - groups did not have many successful readers. A recent study (Hu and Nation, 1999) of reading comprehension, examining adequate vocabulary coverage related to reading for pleasure, found that when 80% of the words of a fiction text (i.e. two unknown words per line) were familiar to readers, none gained adequate comprehension; at 90% to 95% some, but not many,

gained comprehension. It was found that only at about 98% coverage (eight unknown words per 400 word page) was "unassisted comprehension" gained. Among the study's conclusions: "as readability studies show, vocabulary knowledge is a critical component in reading . . . . [though] research so far has not been able to provide a clear guideline about the optimal density of unknown words" (Hu and Nation, 1999, para. 74). While the optimal percent-density of unknown words can not yet be known precisely, Nation has concluded that:

> Knowing academic vocabulary is a high priority for [L2] learners who wish to do academic study in English. After gaining control of the 2,000 high frequency words, learners need to then focus on academic vocabulary. Knowing the 2,000 high frequency words and the Academic Word List will give close to 90% coverage of the running words in most academic texts. When this is supplemented by proper nouns and technical vocabulary, learners will approach the critical 95% coverage threshold needed for reading" (Nation, 1999, p. 274).

Thus, this area of research has been able to establish that the 95% figure is workable as a threshold for known/unknown vocabulary, in EAP study.

The 2,000 "high frequency words" (Nation, using shorthand, is indicating highest frequency headwords/word families), above, are taken from West's GSL, which, though aging, is defended as remaining "the best of the available [general service] lists" (Nation and Waring, 1997, p.13). The AWL is composed of the most frequent word families from a 3,500,000 running word academic corpus containing a balance of the four general fields of science, arts, commerce and law-with each of

these fields further divided into seven subject areas (Coxhead, op. cit.). It is an excellent list because each word contained in the list has a range that includes all 28 (4 × 7) subject areas (the corpus texts are divided equitably across all subject areas), and a frequency in the complete corpus of at least 100. The third list, the technical/other word list can be either imported or composed by the teacher: this list may be designed as field-specific, syllabus-specific, or as a multi-disciplinary list, at the teacher's discretion. In the absence of a technical word list, the concordance program will display all words not occurring in the GSL or AWL in a combined "technical/other" category. A technical word list can be created by the teacher by: (a)selecting words out of the "technical/other" list generated by the program from the source reading text, and then (b)copying into a word processor document all of the technical terms they wish the students to see in a (to be created) customized "technical" word list, and (c)saving this file as a plain text (ASCII) file (with one word per line), naming it "technical", and then (d)placing the file in the "Word List" folder. When the program next runs, it will find and load the newly created list. A breakdown of coverage by the different kinds of vocabulary in a typical academic corpus is seen in Table 1, below.

Table 1. Percent Word-Coverage in the Academic Corpus (Coxhead, op. cit.).

Type of Vocabulary	**% of Coverage**
1st 1000 word families	71.4%
2nd 1000 word families	4.7%
AWL 570 word families	10.0%
Others	13.9%
Total	100.0%

Nation finds that, in terms of EAP/academic coverage, as a rule of thumb (to paraphrase): the most frequent 1,000 GSL words cover about 77%, the second 1,000 cover about 5%, AWL (semi-technical) vocabulary will account for approximately 8.5%-10%, and technical words ("words very closely related to the topic and subject area of the text") typically cover about 5% of the running words of a text (1999, pp.7-13). The total percent coverage then, is 95.5%-97%, which provides a target goal for academic vocabulary acquisition, leading to the unaided comprehension of unknown vocabulary by context, and successful reading comprehension, within a particular academic field. The additional "missing" 5^{+}% represents vocabulary of lower frequency than the first 2,000 GSL words, and words which are semi-technical but not in the AWL list.

We can note the specificity of the AWL to academic texts. The third group of 1,000 GSL word families (words with a frequency of 2,000 to 3,000) covers only about 4.3% of the AWL corpus (Coxhead, 1998; Nation, 1999). Though "there has been little available research comparing the frequency of specific academic words in academic and non-academic texts" (Nation, 1999, p.268), one frequency study of the University Word List (UWL), an 800 word-family academic list, a predecessor to the AWL, showed a coverage of only 1.7% in fiction texts, compared to 10% coverage of academic texts (Nation, 1999). These results indicate the high degree of specificity of academic vocabulary, and suggest that while the typical EFL L2 student will have a goal of comprehending the first 3,000 GSL word families, it is beneficial for the EAP student to focus on semi-technical vocabulary after they have succeeded in comprehending the first 2,000 GSL word families. For the EAP student then, the GSL 2,000 highest-frequency words represent core vocabulary, while the AWL words can be considered high-

frequency vocabulary. Students entering an academic program can be assessed, using a measure such as the Vocabulary Levels Test, to determine whether they have a satisfactory comprehension of the first 2,000 GSL words, and to what extent semi-technical/AWL vocabulary is known.

Generally speaking, acquiring semi-technical vocabulary is challenging for students. Anderson (1980) found that semi-technical words were those most often identified as unknown by learners studying academic texts. By focusing on words within the AWL, students will utilize their study time efficiently, as AWL vocabulary has relevance across academic disciplines – another implication is that these words, studied within the context of academic texts, should provide an efficient and beneficial method of TOEFL preparation (though the relationship between the TOEFL reading corpus and AWL frequency/coverage has yet to be determined).

"VOCAL": Design And Relevance

The primary goal in designing our "multimedia vocabulary concordance and academic lexis (VOCAL)" program is to provide EAP students with options allowing them to quickly discern, from computerized text, words in the three levels of vocabulary; to quickly search for definitions and/or translations of words; to examine collocations from selected corpora; to acquire new vocabulary though VOCALs tripartite, GSL/AWL/Technical 'minimal efficient' approach to vocabulary acquisition. Any selected parts of reading texts, collocations, and word definitions can themselves be combined, organized, and saved. Texts can be analyzed to determine the number and percent-coverage of words in each of the three vocabulary groups existing in that text--this allows teachers (or the student) to "grade" any text for vocabulary

comprehension, and thus organize a graded-reading syllabus, if so desired. Another possibility for students has to do with VOCAL program set-up. Along with the variety of research options built into the program interface, a "Help" section indicates to students (or teachers) how to create micro-corpora. Students with internet access can self-select several internet sites as search defaults, and can also link their own external dictionary software to the program. Thus, students are able to individualize and customize certain VOCAL features, allowing for (learner-centered) participation and involvement in program design and use. Multimedia options include text-to-speech synthesis, and the ability to play audio and video files; e.g., MP3/WAV/AVI performances of source reading texts can be included in the program.

In a recent study conducted at the Prefectural University of Kumamoto, of the four EFL skills, reading was considered the most important by the science faculty (Melton, 2000). The concordance program discussed here is particularly relevant to reading study in the sciences, because (a)reading comprehension is a prioritized goal of EFL education in the sciences, (b)graduate students may be expected to read and potentially contribute to research papers published in international (English language) journals, (c)students utilize computers in other classes and are familiar with computers, (d)computer labs are usually available for classroom use. Details of the features and use of VOCAL will be described in the next section.

Concordance Design And Features

The Main Window

The VOCAL program is primarily operated through its main window

(Figure 1), below:

The "Source Reading Text (SRT)" frame displays the selected reading text. The "Word List" displays words which appear in the SRT. The "Collocations" frame displays collocations, based upon whatever word or phrase, in either the "SRT" or "Word List" frame has been highlighted. Buttons along the bottom of the window provide the user with various 'instant' choices: (a) "Word List" display options (ALPHA=alphabetically, FREQ=by frequency, ASC=in ascending or DSC=descending format); (b)what words are displayed in the "Word List" (ALL=all words, words from the GSL, or AWL, or TECH technical/other list)--e.g., "ALL" displays all words in the SRT in the "Word List" frame, while "AWL" displays only those SRT words that also are contained in the AWL word list; (c) three default external sources can be accessed: two default dictionaries, ENG and TRANS, and a (user-selected) default internet site, (being related to external programs or access nodes, this last group of buttons may be optionally active).

Program operation will typically begin by loading an SRT into the "SRT frame" ("File-open"), and selecting one of the available text files within the SRT folder. SRT texts have been first prepared by a teacher/course designer, either by: (a)scanning and OCR-conversion into a word processor, downloading/copying, or keyboard entry; and then, (b)once a text is in the word processor, it is converted to a plain (ASCII) text format for use in VOCAL, and; (c)placed into the SRT file folder.

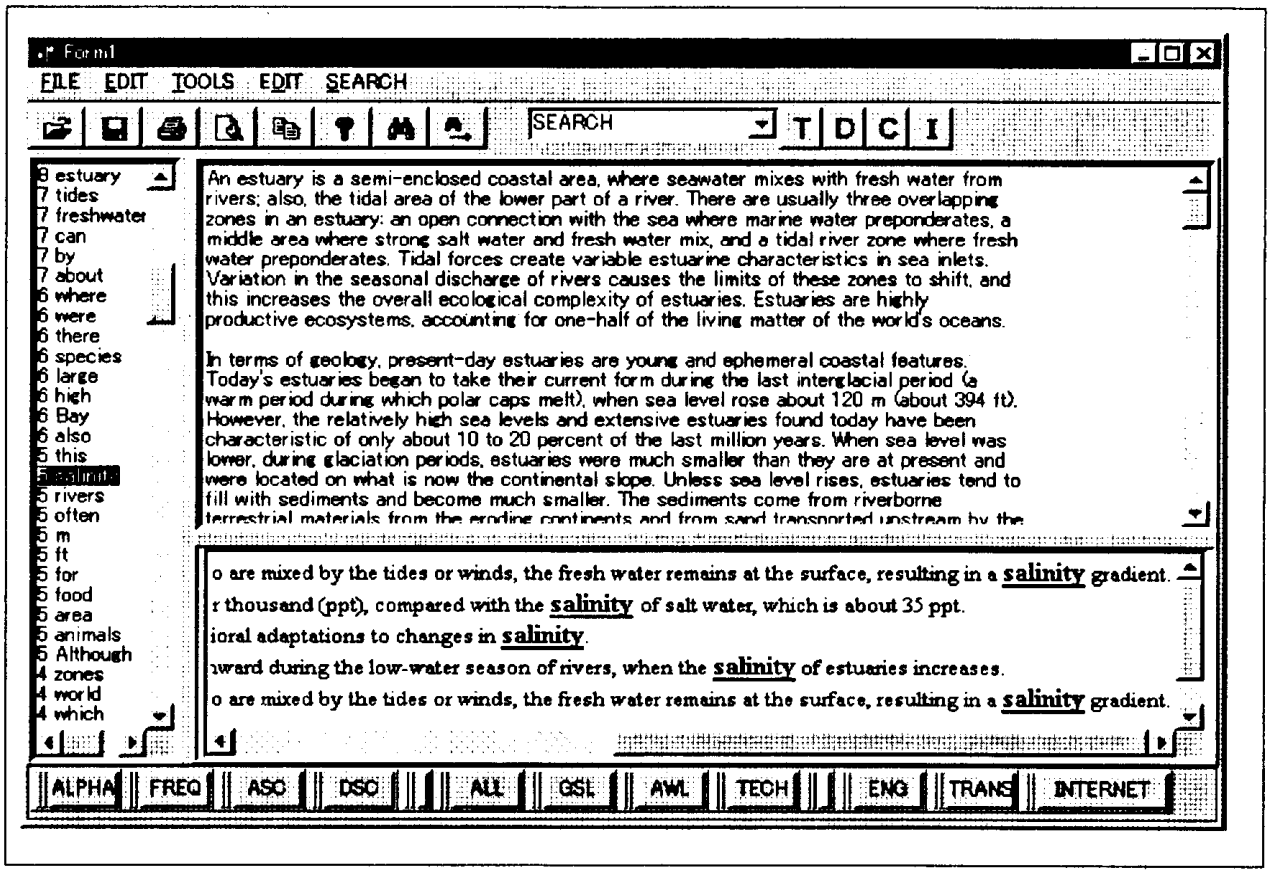

Figure 1. VOCAL program: Main Window

The "Right-Click Mouse" Menu

Further functions are available by right-clicking the mouse. The right-click mouse menu appears below, in Figure 2:

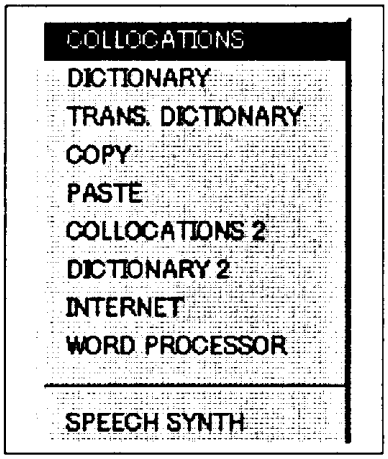

Figure 2. VOCAL program: Right-Click Mouse Menu

(a)"Collocations" finds collocations of user-highlighted sections. Collocations are found by searching the default corpus (all VOCAL defaults are set through the "Tools-set defaults" option); (b)"Dictionary" and (c)"Trans. Dictionary" search these respective external default programs (exactly as the matching buttons at the bottom of the main window); (d)"Copy" and (e)"Paste" functions allow for materials in the program to be saved; (f)"Collocations 2" searches the second default corpus; (g)"Dictionary 2" searches the second default dictionary; (h)"Internet" searches the default internet site; (i)"WORD PROCESSOR" opens up the default word processor for saves; (j)an English text-to-speech synthesizer will 'read and speak' any highlighted word, phrase, sentence, or section.

The above features, taken together, allow a student to quickly find bilingual definitions, collocations, acquire pronunciation information, vocabulary levels, SRT word frequencies, and also organize and save information as files in their word processor.

The "Tools" Menu

Additional features, including multimedia options, default options, and analysis options are found in the "Tools" menu (Figure 3), below:

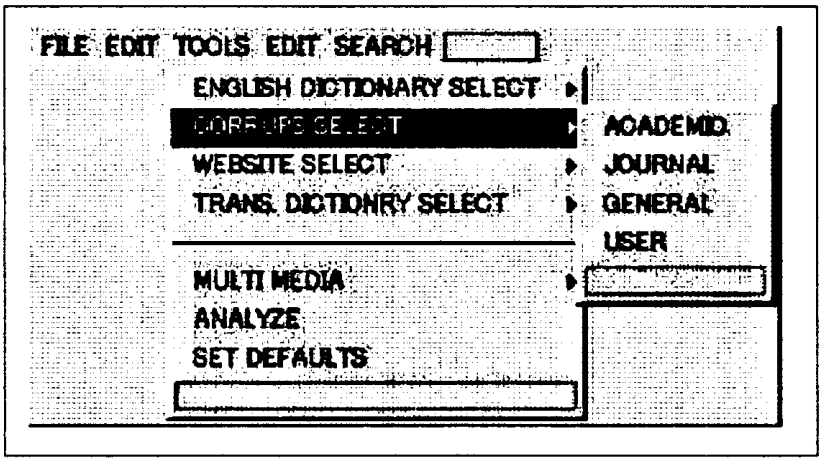

Figure 3. VOCAL program: Tools Menu

The "Tools menu allows for more detailed selection options, as well as certain advanced features. (a)"English Dictionaries" allows for the selection of any one of three defined English dictionaries. For copyright reasons, these dictionaries are external software packages, which will open automatically in a new window when selected; (b)"Corpora" selects any one of up to four separate default corpora: academic, research, journal, general, user. Corpus creation will be discussed in the next section; (c)"Website" allows the selection of three defined websites; (d)"Trans. Dictionaries" selects one of three defined translation dictionaries; (e)"Multimedia" opens a dialogue box containing audio and/or video files, which then can be selected and played. Typically, these would be audio/video recordings of the source texts, able to be listened to/viewed. Multimedia files would be created at the discretion of the teacher, as supplementary educational materials; (f)"Analyze" will analyze the source text and display vocabulary coverage information based upon the word lists contained in the program database, in a fashion similar to the DOS-based "VocProfile" software program (Heatley, Kyongho and Nation, 1997). Coverage information is parsed so that percent-coverage is shown for (a)the 1st and (b)2nd thousand GSL word families, (c)the 10 sub-divisions of the AWL (each division represents 60 word families, in order of decreasing frequency in the AWL corpus), and (d)technical/other words.

Multimedia

Although the focus of VOCAL is written texts, there are several important considerations for the inclusion of multimedia materials. First, sound and visual imagery aid in memory retention and comprehension. Second, the Japanese (or other L2) student of English often has great difficulty with pronunciation, and relating pronunciation with

word spellings. Third, Nation (1999) makes the point that "meaning focused input, language focused learning, meaning focused output, and fluency development – should be seen as opportunities for the development of academic vocabulary knowledge. Thus there should be *listening* and reading activities that encourage the learning of academic vocabulary. . . ." [our italics] (p. 274). A number of interactive learning activities are planned for incorporation into additional program modules – however, it seems appropriate to place multimedia materials directly related to the SRTs within the VOCAL module, thereby allowing for rapid and convenient student access.

The "Search" Entry Box Functions & "File," "Edit," & "Help" Menus

The "Search" window, near the top of the main window, allows the user to enter any word or term, and select a source to search. The four buttons to the right of the search window allow searches in (a)the SRT; (b)an English dictionary; (c)the default corpus (the results will be collocations, appearing in the "Collocations" frame); (d)the default internet search site (this can be a different default than the "internet" select options either at the bottom of the main window, mouse right-click menu, or "Tools" menu). Additional menus include typical administrative software functions: "File" (open, close, print, exit the program), "Edit" ("Copy", "Paste" – there is no 'cut' function, as the concordance is read-only, "Find"), "Help" (program description and control, corpus creation instructions). Excepting for the creation of corpora for collocations, and for the setting of program defaults, the basic features relating to the functioning of the VOCAL program are all accessed through the main window, the mouse right-click menu, or the "Tools" menu, as detailed above.

Saving & Organizing Information

Information, such as terms, definitions, collocations, and notes, can be easily organized by students in their word processor, by creating a template or default table, such as the sample illustrated below, in Table 2:

Table 2. A sample expandable table for organizing and saving VOCAL information.

TERM	JAPANESE	DEFINITION	NOTES
SRT EXAMPLE &/or COLLOCATIONS			

All of the terms a student wishes to save from within a SRT can be included within one document, and given the same title as the SRT. Word-processor documents can be printed out and studied anywhere, as can SRTs, if they are of reasonable length. Thus, VOCAL can provide an organizational study structure to aid in memorization and course review. As VOCAL can be distributed in CD-ROM format or downloaded, a student can conceivably retain the VOCAL program, the SRTs, and their own annotations, into the foreseeable future in an organized, accessible format.

Collocations: Creating "micro" corpora

Corpus Sets

The essential goal of providing collocations for students is to demonstrate a number of "expert performances" of usage within the target-

field(s) or genres of learning. Students "almost certainly need many corpora rather than one . . . [they] need a rich set of potential models for their own language behavior . . . the most useful corpus for learners of English is one which offers a collection of expert performances" (Tribble, 1997, paras. 8 - 11). In order to provide both field - specific collocations, and also a wider circle of collocations related to EAP study, the VOCAL program calls for at least one field - specific corpus, and indicates the possibility of assigning up to four different corpora in total for use in the program: "academic," "journal," "general," and "user" (definition tags are located in the "Tools" menu). These four corpora are defined as: (a)the academic corpus—representing a corpus containing field - related and field - specific writings, essays, texts: this corpus demonstrates 'performances' oriented to the educated reader in the field; (b)the journal corpus—representing a corpus containing published research papers: this corpus represents 'performances' at the highest professional level of the field; (c)the general corpus—representing a larger, multi - field corpus containing 'performances' across, for instance, all scientific fields, or, taking a larger view, across multi - disciplinary fields; (d) a "user" corpus definition allowing user to create and define a corpus.

These four defined selections can be customized (though the menu tag - names cannot be altered). Any ASCII text file, of whatever content and composition, can however be treated as a corpus, by linking it to one of the corpus definition - tag names, if so desired. Additionally, more than four sets of corpora can be kept within the corpora folder, and rotated — though only four corpus sets can be defined by VOCAL (through the "Tools" definition - tag names) at one time.

Do-It-Yourself: How To Create "Micro" Corpora

While it is always possible to import and use a corpus from elsewhere (as long as the corpus is in ASCII text format, and contains no headers, keywords, etc.), there are often commercial, technical, and legal issues which prevent the use of purchased corpora by language teachers. There are, however, methods for creating non-standard micro-corpora — collections of expert performances in genres which have relevance to the needs and interests of EAP learners. Such micro-corpora have been defended as being as useful as a larger "professional" corpus, and in some cases preferable, for students who are learning how to "[read and] write formal, professionally oriented texts" (Tribble, op. cit., para. 13):

> I have been able to construct themed, twenty to thirty-thousand word micro-corpora in fifteen to twenty minutes. Although such a corpus sounds insignificantly tiny when compared with the huge corpora which can now be accessed, I would argue that if one wishes to investigate the lexis of a particular content domain (e.g. health) a specialist micro-corpus can often be more useful than a much larger general corpus. For example in the written component of the BNC Sampler (1,000,000 words) there are no instances of "cancers". An Encarta micro-corpus of health articles (24,805 words) gives 33 usefully contextualized examples" (Tribble, op.cit., para. 24).

A 'quick and dirty' source for both the "academic" corpus and a "general" EAP corpus is the multimedia encyclopedia, e.g. the Microsoft *Encarta* CD-ROM Encyclopedia (Microsoft, 2000), as mentioned above. (Other multimedia CD-ROM encyclopedias include, for

instance, Hutchinson's and Grolier's). A few simple steps lead to the creation of a corpus of 35,000 to 50,000 running words (30 - 40 articles), in about an hour's time. The steps using *Encarta* are:

1. Open a word processor.

2. Use the Encarta search engine to locate texts: e.g., a search using the term "ecology" yields a number of sub - headings from which articles useful for environmental science classes can be culled: activists, adaptation, animal health issues, atmosphere and weather, balance in ecology, biodiversity, biosphere, biomes, ecosystems, ecology of specific regions, ecology, patron saint of, environmental damage, environmental protection, events involving ecological issues, evolution, theory of, groundwater, human ecology, literature, as a subject in, longshore drift, natural disasters, politics, research and study, etc.

3. After locating an *Encarta* text, select "copy" from the menu and copy/paste into a blank, open word document.

4. Edit to remove any headings, and change the article lead - in. e.g., the article "Estuary" begins: "I. Introduction / / Estuary, semi - enclosed costal area, where seawater mixes with fresh water . . ." (Encarta, op.cit.). To edit, remove: "I. Introduction," and change the first line to read: "An estuary is a semi - enclosed . . ."

5. Save as a plain text (ASCII) file, and/or add more texts. (If one prefers, one can save as a plain text file first, and then edit the text, thus reversing steps #4 and #5.)

Currently, *Encarta* contains 42,000 articles (written by 5,000+ authors). Article size ranges from approximately 500 to as long as one (seven-part) 50,000 word article; the articles are fairly evenly balanced over nine meta-subject areas. (There are over 4,000 articles under the heading, "Physical science," and nearly 5,000 under the heading "Life science.") Assuming an average article to be around 1,200 words, this yields a total possible corpus of 50,400,000 running words. The "Estuary" article, quoted above (1,800+ words), selected from the keywords "ecology − biomes − estuary" and chosen at random, was contributed by Michael Goulding, Ph.D., Director, Amazon Rivers Project, Rainforest Alliance. It is representative of an authentic 'expert performance' of formal, professionally-oriented writing, and of potential interest to L2 students engaged in environmental science study. (The full article can be accessed from the worldwide web at: http://encarta.msn.com "estuary".) A teacher or student is easily capable of creating numerous sets of subject or field-oriented corpora, using a CD-ROM encyclopedia, for use in searching for EAP-genre collocations in the VOCAL program. (Articles such as these might also make excellent SRTs!) To create a "general" corpus one would use the same text resources, but seek to construct a balanced corpus, spanning several fields.

A third corpus, the "journal" corpus can be constructed from international journals in the appropriate field(s) of study, by a process of: (a)scanning; (b)use of optical character recognition (OCR) software; and then (c)conversion to ASCII text files, as detailed above, being careful to eliminate headings, tables, references, and inserts. Assembling a journal corpus is thus a more time-consuming affair, unless journal articles are available on CD-ROM or online−which is increasingly becoming the case. Again, it is up to the corpus designer how

broadly or narrowly the journal corpus is defined. It should be mentioned that though assembling a scanned-in "journal" corpus, or any other scanned corpus, is labor-intensive, it need be accomplished only one time to serve for a number of years or decades as a viable educational resource. As well, any of the corpora mentioned above can be expanded incrementally.

Discussion

To the present, computer-based concordance software has not been designed with ESP parameters in mind, and would be difficult to integrate into the ESP classroom. Particularly, existing concordances are most relevant for linguistic investigations, rather than ESP-oriented reading, which is oriented towards the extraction of information. One of the principles defining the purpose of reading in ESP study as opposed to general EFL study, is the functional shift from "text as a linguistic object (TALO)" to "text as a vehicle for information (TAVI)" (Dudley-Evans and St. John, op. cit.; Johns and Davies, 1983). "The key principles [are] that, for ESP learners, extracting information accurately and quickly is more significant than language details; that understanding the macrostructure comes before language study; and that application of the information in the text is of paramount importance" (Dudley-Evans and St. John, op. cit.; p. 96). The few software packages that are relatively accessible in terms of use and price (e.g., MonoConc Pro from Athelstan Press, and Wordsmith Tools from Oxford University Press), do not include word lists, dictionary lookup features, multimedia integration, internet connectivity, or possibilities for the 'instant' selection of different corpora for collocations, when view-

ing a SRT. The 'target user' of these packages is mainly the linguistics researcher, or student. EAP science reading classes in the university do not represent a 'pure' ESP setting – nevertheless, the "TALO to TAVI" principle applies, and VOCAL design is oriented to an ESP/EAP "TAVI" approach to an interactive, autonomous, learner-centered reading syllabus.

Two of the main drawbacks of any CALL educational tool are the necessity for a computer lab, and the inherent lack of portability. The VOCAL program particularly can be critiqued in terms of the set-up effort involved in order to create a "complete" educational tool, according to the intended design. Issues include the setting up of SRTs, corpora, and the technical word list. As well, VOCAL relies on external software programs, which may need to be purchased, e.g, the dictionary software. On the plus side, most science faculties with computer labs and EAP programs already possess the needed software; in cases where purchase is required, the cost is low. As well, the effort involved in creating SRTs and corpora will result in organized, permanent, and revisable course materials.

One or two reading classes per week cannot provide enough study time for the motivated student to make rapid progress in vocabulary comprehension. By creating a multimedia CALL concordance and collocation educational tool which incorporates a lexis acquisition strategy into its design, a student can study autonomously, to advantage. Students are also able to easily organize, customize, and save whatever information thy deem relevant. As well, VOCAL is useful for advanced L2 learners of English who wish to develop reading/writing skills at the highest professional level – for the post-graduate student or professor who desires to both read and publish in international journals.

This paper has provided a rationale for the design of an educational

tool with a specific orientation and goal: accelerated study in science reading and vocabulary comprehension for L2 university students. Currently, a 'proof of concept' prototype is being developed. It is hoped that in the following months the VOCAL program will be available for pilot studies in the classroom.

References

Anderson, J. (1980). The lexical difficulties of English medical discourse for Egyptian medical students. *English for Specific Purposes (Oregon State University) 37* (4), 23 - 35.

Berge, Z., & Collins, M. (1995a). *Computer - mediated communication and the online classroom, vol. 3.* Victoria, Australia: Hampton Press.

Berge, Z., & Collins, M. (1995b). Computer - Mediated Communication and the Online Classroom in Distance Learning. *Computer - mediated communication magazine* 2 (4), 6 - 10.

Coxhead, A. (1998) An academic word list. *E.L.I. Occasional Publication Number 18*, LALS, Victoria University of Wellington, New Zealand.

Dudley - Evans, T. & St. John, M. (1998). *Developments in English for special purposes: A multidisciplinary approach.* Cambridge, UK: Cambridge University Press.

Encarta Encyclopedia Deluxe 2000. (2000). Multimedia CD - ROM Encyclopedia. Seattle, WA: Microsoft Corp.

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics 11* (4), 341 - 363.

Heatley, A., Kyongho, H. & Nation, P. (1997). *VocabProfile*. DOS - based Vocabulary Analysis Software. English Language Institute, Victoria University, P.O. Box 600, Wellington, New Zealand. Available on the worldwide web at: http://www.vuw.ac.nz/lals/

Hirsh, D. and Nation, P. (1992) What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language 8* (2), 689 - 696.

Hutchinson, T., & Waters, A. (1987). *English for specific purposes*. Cambridge, UK: Cambridge University Press.

Hu, M. & Nation, P. (1999). *Unknown vocabulary density and reading comprehension*. In progress. Victoria University of Wellington, New Zealand.

Johns, T. and Davies, F. (1983). Text as a vehicle for information: The classroom use of written texts in teaching reading in a foreign language. *Reading in a foreign language 1* (1), 1 - 19.

Johnson, J. O. (1999). Collaborative design, constructivist learning, information technology immersion, & electronic communities: A case study. *Interpersonal computing and technology: An electronic journal for the 21st century*, (7:1 - 2), available at: http: //emoderators.com/ipct - j/1999/n1 - 2/alexander.html

Kikue, K. & Iiyoshi, T. (1996). *Maruchimedia dezain ron* [Art of Multimedia: Design and Development of the Multimedia Human Body]. Tokyo: ASCII Co.

Klinmanee, N. and Sopprasong, L. (1997) Bridging the vocabulary gap between secondary school and universtiy: a Thai case study. *Guidelines 19* (1), 1 - 10.

Kyongho, H. & Nation, P. (1989) Reducing the vocabulary load and encouraging vocabulary learning through reading newspapers. *Reading in a Foreign Language 6* (1), 323 - 335.

Kyongho, H. & Nation, P. (1995) Where should general service vocabulary stop and special purposes vocabulary begin? *System 23* (1), 35 - 41.

Laufer, B. (1992) How much lexis is necessary for reading comprehension? In P.J.L. Arnaud and H. Bejoint (Eds.) *Vocabulary and applied linguistics*. Macmillan, London, 126 - 132.

Laufer, B. (1994) The lexical profile of second language writing. Does it change over time? *RELC Journal 25* (2), 21 - 33.

Laufer, B. and Nation, P. (1995) Lexical richness in L2 written production: Can it be measured? *Applied Linguistics 16* (3), 307 - 322.

Laufer, B. and Nation, I.S.P. (1999) A vocabulary size test of controlled productive ability. *Language Testing 16* (1), 36 - 55.

Liu Na and Nation, I.S.P. (1985) Factors affecting guessing vocabulary in context. *RELC Journal* 16 (1), 33 - 42.

McClintock, R. (1992). *Power and pedagogy: Transforming education through information technology*. New York: Institute for Learning Technologies, 1992.

Melton, J. (2000). Preparing materials for English for specific pur-

poses: A faculty-wide needs analysis. *Language Issues: Journal of the Foreign Language Education Center 6* (1), 11-30.

Mieszkowski, K. (2000). Microsoft's brave new dot-net world. *Salon.com*. Accessed June 23, 2000. Available on the worldwide web at www.salon.com.

Nation, P. (1999) *Learning vocabulary in another language*. E.L.I. Occasional Publication Number 19, LALS, Victoria University of Wellington, New Zealand.

Nation, P. (1997). The language learning benefits of extensive reading. *The Language Teacher 21* (5), 13-16.

Nation, P. (1993). Vocabulary size, growth and use. In R. Schreuder and B. Weltens (eds) *The bilingual lexicon*. Amsterdam/Philadelphia: John Benjamins, 115-134.

Nation, P. (1990). *Teaching and Learning Vocabulary*. Rowley, Mass: Newbury House.

Nation, P. (1983). Testing and teaching vocabulary. *Guidelines 5* (1), 12-25.

Nation, P. & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt and M. McCarthy (eds) *Vocabulary: Description, acquisition and pedagogy*. Cambridge, UK: Cambridge Press. 6-19.

Nation, P. & Worthington, D. (1996) Using texts to sequence the introduction of new vocabulary in an EAP course. *RELC Journal 27* (2), 1-11.

Read, J. (1997). Vocabulary and testing. In N. Schmitt and M. McCarthy (eds) *Vocabulary: Description, acquisition and pedagogy*. Cambridge, UK: Cambridge Press. 303 - 320.

Soo, K. (1998) Control and CALL Software Design. *CAELL Journal 8* (3), 8 - 14.

Stahl, S. (1990). Beyond the instrumentalist hypothesis: Some relationships between word meanings and comprehension. Technical report no. 505 of the Center for the Study of Reading, University of Illinois at Urbana - Champaign.

Steinberg, E.R. (1989). Cognition And Learner Control: A Literature Review. *Journal Of Computer - Based Instruction 16* (4), 117 - 121.

Sutarsyah, C., Nation, P. and Kennedy, G. (1994) How useful is EAP vocabulary for ESP? A corpus based case study. *RELC Journal 25* (2), 34 - 50.

Tribble C (1997) Improvising corpora for ELT: Quick - and - dirty ways of developing corpora for language teaching. In Melia, J. & Lewandowska - Tomaszczyk, B. (ed.) *PALC '97 Proceedings*. Lodz, Poland: Lodz University Press.

West, M. (1955). *Learning to read a foreign language*. London: Longman. 2nd ed.

West, M. (1953). *A general service list of English words*. London: Longman, Green & Co.

Wodinsky, M. and Nation, P. (1988) Learning from graded readers. *Reading in a Foreign Language 5* (1), 155 - 161.

Wood, A. (1997). International scientific English: Some thoughts on science, language and ownership. *Science Tribune 2* (4). Available on the worldwide web at: http://www.tribunes.com/ tribune/ art97/wooda.htm.

Zechmeister, E., Chronis, A., Cull, W., D'Anna, C., & Healy, N. (1995). Growth of a functionally important lexicon. *Journal of Reading Behavior 27* (2), 201 - 212.