# Classroom Concordancing: Vocabulary Materials Derived From Relevant, Authentic Text

### **Vance Stevens**

Abstract — Though concordancing is widely acknowledged to be a tool of some potential in the hands of second language learners, there is little, if anything, in the literature to suggest how such a tool can be introduced to learners unfamiliar with computers and/or unschooled in basic research methods. This article demonstrates how concordances can be used to prepare classroom vocabulary materials which are both authentic and, assuming the databases used are ones pertinent to the students, relevant. It also shows how concordance—based exercises can be used to enhance student competence in both semantic and syntactic elements of the language they are studying.

### Introduction

Concordancing has recently become increasingly associated with computer-assisted language learning (CALL); for example, concordancing figured prominently in Higgins and Johns (1984), and Goethals (1987) describes how concordances were used to generate a variety of exercise types (see also, Stevens, 1988a). Due to its ability to promote authenticity of script, purpose, and activity, Tim Johns (1989a; p. 15) calls the concordancer "one of the most powerful tools that we can offer the language learner, as it has been for the past quarter of a century for researchers in fields as diverse as lexicography, stylistics, and descriptive syntax." However, almost all references to concordancing in the CALL literature refer to applied linguistics applications, with implementations pertaining to students learning to speak a language limited to those cases where such students act with the maturity and sophistication of applied linguists.

Johns' students are a case in point. In a recent report on concordancing, Johns' (1989b) subjects are overseas postgraduates, while in his earlier article (1989a), he attributes his successes with concordancing in part to the fact that his students are "familiar with research methods across a number of different disciplines, and are used to looking for underlying patterns and regularities in data that on first inspection appear confused and irregular." Concordancing would logically appeal to such students, giving as it does the capacity to efficiently search and extract sufficient linguistic data to resolve the kinds of questions that inquiring minds should be asking about a language under study.

But it cannot be assumed that all language learners are accustomed to

Address correspondence to: Vance Stevens, Language Centre, Sultan Qaboos University, Box 32493, Al-Khod, Sultanate of OMAN.

independent inquiry. Indeed, there is a large subset of language learners who through cultural influences or academic immaturity cannot be expected to search automatically for patterns in a welter of linguistic data. Can such students benefit from concordancing? This paper is based on the premise that they can, if given appropriate guidance. Specifically, this paper reports on a method of introducing concordancing to undergraduate Omani male and female science students at the English-medium university in their country. In addition, it shows how concordancing enables the creation of vocabulary exercises which, rather than being artificially contrived, are based in authentic text directly relevant to what the students are studying in their various courses.

#### Method

There are 10 IBM-XT computers in the CALL lab in the Student Resource Centre (SRC) run by the Language Centre at Sultan Qaboos University (SQU). On the hard disk of each computer is a language database currently comprising several megabytes of text, and growing steadily. The texts are soft copy versions of the biology, geology, physics, and chemistry textbooks that first-year students in the College of Science have used since 1987 at SQU. Departments in the College of Science produce the textbooks on their word processors and then pass copies of the disks on to the Language Centre. Also, some of the texts are from textbooks produced in the Language Centre itself. Thus, the texts represent most of the reading required of first-year students in the College of Science at SQU, except that they have in some cases been slightly adapted for presentation in the computer medium; for example, for compatibility with our text manipulation programs, formulae and foreign characters have been either removed or incorporated into the prose where they had originally been set off on separate lines. In addition, the database includes spoken English in the form of more than two dozen first-year chemistry, biology, physics, and geology lectures transcribed by Chris Arden-Close and Roger Griffiths in the course of their research at the Language

Students and teachers can access these databases by using a concordance program developed in-house by Dave Poulton. The concordance program finds each instance of where a specified string is used in the text selected and displays the context (79 characters default) on the screen with the concordanced string appearing neatly in the middle of each sample. The concordancer works with greatest flexibility from the DOS command line, but for students to use it from DOS requires a more fundamental understanding of computing than can be imparted to students or their teachers in the time allotted. Therefore, the concordancer is adapted to student use by means of an interface program designed by this author. The interface allows the students to boot the program from a floppy disk (i.e., place the disk in drive A and turn the computer on). A batch file automatically accesses the database stored on hard disk and runs the interface program, which displays available texts and allows their selection via cursor movement keys. Furthermore, the program allows the previously

selected search string to be used with any of the texts so that no retyping is necessary to carry a particular search over to additional texts.

For teachers or students who have more than a rudimentary knowledge of computers, greater control can be gained over concordance output by operating the program from DOS. In this case, the context can be varied up to 255 characters. Of greatest use to materials developers, output can be directed to disk files for manipulation with a word processor later.

### SQU Language Centre Use of Concordance Facilities

Students can use the concordance program to find out how words are used in the texts in which they actually encounter them. In practice, when students run concordances on words they find in vocabulary lists in their science textbooks or in the accompanying English support texts, they frequently discover that the word appears in the text only once or—and this is all too often the case—not at all. Students are provided such lists ostensibly because the words carry over to other reading they must do; yet, concordances reveal that many of the words in their lists simply do not recur, and sometimes don't appear at all in the textbook in which the word is listed as if it were crucial vocabulary.

It should not be left to students to test the frequency of occurrence of words on their vocabulary lists; hence, this is a prime area where concordance programs can assist teachers, especially those involved in materials writing. Given access to a corpus of texts that students at the Language Centre are expected to read while they are taking Language Centre courses (which are designed to support the courses for which the textbooks were written), it is now possible for materials writers to concordance words in vocabulary lists to ascertain whether they actually appear in the textbooks, and to deal appropriately with words occurring rarely or not at all.

Concordances make it possible for materials writers to gain insights into many other features of those texts. For example, one Language Centre coordinator requested a concordance on some of the discrete linguistic items taught in the English textbook used in his course and discovered that many of those items never appeared in the science textbooks the students were reading. As a result, the English textbook was replaced by materials more appropriate to the students' coursework outside of English class.

Finally, concordances make possible the creation of vocabulary exercises based in authentic text relevant to the students. Contexts for vocabulary practice can be drawn with minimal effort from the textbooks that the students are reading (or have previously read) in their other courses. Also, exercises can be created having advantages over traditional exercise types. For example, a traditional vocabulary exercise might present a paragraph from which several words have been extracted and arrayed at the top of the page, and the students must replace the words in the blanks they came from. The problem with this kind of exercise is that when students place one word incorrectly, they are likely to compound the error because that word is no longer available for

Here are some words used in a biology lecture during the 5th week of a first-year biology class:

Classification	encangle	COII
classification	entangle	coil
certain	hierarchy	oral
attached	extend	obtain
anchor	excess	obstacles

Below, you find the result of a "concordance" made on some of these words. In this concordance, a computer looked at all the readings in the first-year biology workbook. Then the computer printed each line containing those words. (The computer doesn't know where words or sentences begin or end; it just prints the line.)

DIRECTIONS: Replace each BLOCK of blank spaces below with ONE WORD from the word list above

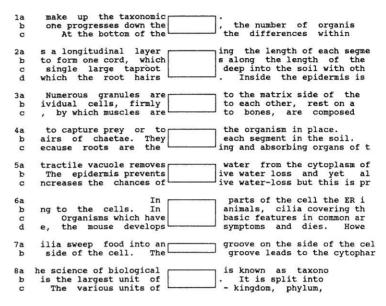


Figure 1. Concordance vocabulary exercise, biology.

placement in its correct blank. A domino effect is set up resulting in low performance on such exercises, causing students to feel inadequate and frustrated.

An improvement on this exercise type is made possible through imaginative use of concordancing, as shown in Figure 1. The program is run from DOS, and output is saved to disk files. From these files, the most revealing contexts for the same word can be selected for the exercise. With a word processor, the search word itself is then blanked out of each context. Since in the concordance output, the concordanced word is always centered within its context, the blanks in the set of contexts will always occur one directly under the other, so that the blank becomes a box several lines deep. A number of these sets of contexts can be presented and the words which go in the boxes can be arrayed at the top of the page, as in the traditional exercise.

The latter exercise is as efficacious as the former, in that it still reinforces the meaning and usage of the words in question, but the multiplicity of contexts, itself a desirable characteristic, reduces the likelihood of error with concomitant gains in student performance and self-confidence. Furthermore, the fact that the contexts are truncated at each end can be put to further advantage if students are requested to fill in words immediately preceding and following the context fragment and encouraged to guess at what the entire sentence might have been. Finally, if the contexts themselves are taken from materials the students are studying in other courses, then the topics and the language in the texts will be familiar and relevant, increasing student motivation to study such exercises as well as expanding opportunities for exploiting them.

Tom Cobb (also at Sultan Qaboos University), working in Macintosh Hypercard, has integrated this concept into a battery of text-manipulation programs that operate from bodies of text stored in the computer and arranged by category. Concordances can be created from one or more individually selected texts, all texts in a category, or all texts in the computer. The programs use the concordances in two ways. One of these is by creating exercises in the format described above, in which students are shown successive samples of concordance output (with the search—word blanked out) and asked to identify the search—word from among distractors selected randomly from the text. The other way is as a vocabulary help option. In this case, students can arrive at the meaning of any word they "point" to by requesting a concordance on that word. They can also request concordances of closed-out words; though those words in the concordance output will also be blanked out. Students then have extra contexts for discovering the mystery word.

The foregoing are examples of how concordances are not only of direct use to language learners, but can be indirectly beneficial when they allow teachers and materials developers to gain insights into the language in textbooks that the students are studying concurrently with their language courses, and to produce authentic vocabulary exercises derived directly from those texts. We now examine how concordance—based vocabulary materials can be used to train students to run exploratory concordances on their own.

### Creating Classroom Vocabulary Materials From Concordance Output

Materials developers with access to concordances will face two problems, especially with students who have little experience with either experiment—based learning or computer-based technologies. Both problems have already been mentioned. The first is that students will not know how to approach a stream of linguistic data with the view to elucidating patterns in the language under study. The second is that students may be familiar with none of the basic concepts of computing that program developers may taken for granted. Besides the usual problems dealing with machines, students may have no conception of how the data can possibly derive from sources they should be interested in. To compound the problem, their teachers may have no more knowledge of computing than they do. Therefore, it probably helps both

students and teachers if the first exposure to concordancing takes place in the classroom and on paper. Exercises such as that shown in Figure 1 present students with their first glimpse of concordance output. Since the purpose is to encourage students to produce their own concordances, it is important at this stage that the students understand how such exercises are made, and in particular what relationship, if any, the text database has with their other coursework. These exercises also help teachers understand the process.

The creation of such exercises is largely mechanical and can be done by anyone who can operate a word processor. It is no more time consuming and is less demanding than creating the kind of exercise in which one tries to construct a wall of artificial text around a handful of vocabulary items. Whether or not the corpus is drawn from familiar texts, the students appreciate the fact that they are dealing with authentic language and all its unpredictable insights. Furthermore, because they have more contexts to work with, hence double checks on their work, students succeed more often with these types exercises than with the traditional ones, and this bolsters confidence and feelings of accomplishment.

The concordance output can also be used as a lesson about the purely mechanical way in which computers work. One would think this is obvious from the way in which words and sentences are arbitrarily truncated (typical concordance programs print a set number of characters to a line); however, some students and teachers initially assume there is a fault either with the program or with the machines because there appears to be "something wrong" with the concordance output (Stevens, 1988b). Arbitrary truncation of the beginnings and endings of lines is an aspect of concordancing that the fastidious tend to find objectionable. However, it can be used to advantage, as in the following exercise type.

Figure 2 shows an exercise in which students are presented with the same samples of concordance output as in Figure 1; but here, attention is drawn to line endings and beginnings, and students are asked to fill in what missing information they can. What is missing is more obvious for some blanks than for others. Where the corpus is drawn from textbooks the students use, the discourse may be more familiar to the students than they are to their English teachers, in which case the students can often use recall to fill in blanks that defy teacher intuition.

One of Johns' (1989a) principles underpinning the use of concordances with students is that teachers are most effective when they are most at risk. In this exercise type, teachers are highly exposed, since they will rarely know the one correct answer, and may be at as great a loss as the students to guess plausible responses to each blank. But there is scope here for good teachers to work with, as opposed to at the head of, their students, and to guide student experimentation in pursuit of discovery rather than acting as the sole source of knowledge in the class.

This exercise type is a good one for developing a sense of predictability in language based on syntactic and semantic clues. Some knowledge of the subject matter is needed to resolve many of the truncations satisfactorily, but

This exercise will help you discover how much you know about sentences you see in concordances even though the sentences are not complete DIRECTIONS: Fill in the blanks with the missing WORDS or LETTERS. Make your best guess there may be more than one possible answer. one progresses down the hierarchy, the number of organis single large taproot extends deep into the soil with oth root hairs extend. Inside the epidermis is which 3a Numerous granules are attached to the matrix side of the b \_\_\_ividual cells, firmly attached to each other, rest on a \_ . by which muscles are attached to bones, are composed airs of chaetae. They anchor each segment in the soil. ecause roots are the anchoring and absorbing organs of t\_ tractile vacuole removes excess water from the cytoplasm of b The epidermis prevents excessive water loss and yet al c \_\_ncreases the chances of excessive water-loss but this is pr\_\_

Figure 2. Follow-up concordance vocabulary exercise, biology.

teachers can facilitate the process by making sure students are attuned to the linguistic clues. For example, in Figure 1, the blank preceding sentence 1b must contain the word as, and the word truncated at the end is obviously organisms. Students who recognize this sentence as coming from a unit on taxonomy recall further that the number of organisms in any one taxon decreases as one progresses down the hierarchy, so additional words are supplied by extrapolation. Similarly, a discussion of sentence 3c not only highlights the syntactic constraint on the preposition following the word composed, but leads to a discussion of tendons and their biochemical make-up. Teachers could lead students to the word tendon by teasing out the relative clause construction, "muscles are attached to bones by...." Similarly, in sentence 2d, the relative pronoun which ought to follow the noun it replaces; therefore, the missing word should be the thing the root hairs extend, that is, the epidermis, or absorptive surface area. Thus this type of exercise is not only rich in opportunities for discussing grammar, but it can lead to open-ended extension of the topics.

The next step in preparing students to concordance on their own is to show them how the concordances they have been working with were prepared; however, there are two difficulties to overcome. First, students must understand the concept of a *string*, as opposed to a *word*, since the computer searches for strings and has no knowledge of words beyond the fact that words are likely to be bounded by spaces. It is specification of a string followed and/or preceded by a *wild card* that creates the illusion that the computer extracts from its database every instance of all possible derivations of a word. The second problem for language learners then is to anticipate how much of the root

must precede the wild card in order to extract from the database as many derivations as possible of the word under consideration without simultaneously collecting distractingly irrelevant data.

For example, a student asking for a concordance of the exact string diffuse will miss all occurrences of the words diffused and diffuses. The student could get a concordance of all three words by stipulating the string diffuse, where the tilde is a wild card; that is, the student has asked the concordancer to present every instance in the database where the eight characters, space, d, i, f, f, u, s, and e, occur together and in that order, followed by any other character or set of characters up to the next space, or word boundary. The concordancer will then present all instances of the desired words in their contexts, but will not display any instances of the words diffusion or diffusing. If the student wishes to play it safe and asked for a concordance on the string diff, then contexts containing different, difficult, differential, and so on would be added to the data. Obviously, the correct string to ask for is diffus, but this knowledge comes to us in hindsight, or, as proficient speakers of English, from our knowledge that words such as diffusion and diffusing exist and must be taken into account. Learners of English, on the other hand, are unable to make equally accurate predictions, and this, along with spelling, causes them their greatest difficulties when they run their own concordances.

The only real solution to this dilemma is experience with the language accompanied by extensive practice with word derivations. As a short cut, students can at least be shown how the concordances for the exercises they have been using were derived. This can be done by means of handouts in class. As follow-up, the students can then be asked to give the string and wild card combinations that elicited certain sample concordances. An example of the latter exercise is given in Figure 3.

### Teaching Students to Run Their Own Concordances

After the initial exposure to concordance-derived vocabulary exercises, students can be sent to the computers for hands-on practice. Before they start, they will need to know essentially what is in the text files; a handout or wall chart can give them this information in a way they can reference it as needed. They then need to know exactly how to use the program to access individual or concatenated text files.

Using the concordance program should be as simple as possible. Considerations in the interface should include stopping the listing in the event of screen overflow as well as dealing with cases where the program is unable to find any instances of the string requested because the corresponding word isn't in the database, or has been misspelled. With large databases, the computer may take some time while appearing to produce no result, so students should be forewarned to expect this, and the program should be set up to deal benignly with random keypresses made by impatient students during periods of apparent inactivity.

Worksheets should be designed to walk students through the program

This exercise will teach you how to use a TILDE [~] to get concordances of all the FORMS of a word you are looking for.

DEFINITION: a STRING is a group of letters taken together. A string can be a word, but it doesn't have to be a word

EXAMPLE: The string: anchor~ was used to get the following sentence segments from the text file BIO-EUS.TXT. Notice how the tilde got 'anchoring' as well as 'anchor'.

to capture prey or to anchor the organism in place.
airs of chaetae. They anchor each segment in the soil.
ecause roots are the anchoring and absorbing organs of t

DIRECTIONS: In the blank spaces, write the string that you think extracted the following texts. Don't forget to add the tilde:

ortant to be able to classify them. The science of biol he science of biological classification is known as taxono is the largest unit of classification. It is split into The various units of classification - kingdom, phylum,

tractile vacuole removes excess water from the cytoplasm of The epidermis prevents excessive water loss and yet al ncreases the chances of excessive water-loss but this is pr

s a longitudinal layer extending the length of each segme to form one cord, which extends along the length of the single large taproot extends deep into the soil with oth which the root hairs extend. Inside the epidermis is

r ER. The cavities are connected with each other, and t ction is performed by connective tissue which binds orga from its function that connective tissue must be strong most fundamental type of connective tissue is areolar tissue y: beneath the skin, connecting organs together, and

Figure 3. Follow-up concordance vocabulary exercise, biology.

operation as closely as possible, then gradually release them to their own experimentation. Figure 4 gives an example of such an exercise. The "instructions on page 1" are those for starting the interface program.

Although obvious, it is too important not to mention that materials developers should scrupulously follow their own instructions to ensure that the concordances do indeed turn up data in manageable quantities. Students easily become frustrated when their first attempts at a concordancing produce no or unexpected results. On the other hand, some unexpected results can be worth exploiting. For example, a concordance of combin in our database turns up an instance of the word combing, at which point students are asked to find the word that is out of place among the various forms of the word combine, and to speculate about why this word should have appeared at all. The answer reinforces the concept that the computer searches arbitrarily for strings, not for semantically related words.

Students can be led to experimentation through the device of chart completion, in which parts of the chart are gradually withdrawn. At first, they Seeing how the program works --

What

a. Follow steps 1 and 2 in the instructions on page 1.

Choose the text file CHEM1989.TXT.

b. At step 3, type this word: combin~
c. Continue the concordance by doing step 4.

You should be concordancing combin~ in CHEM1989.TXT

1 4 2. 5.	orms of "combine" do you find?	fferent forms of
2. 5,	4	N
7.1. <del></del>	5	
3 6	6.	

Figure 4. Exercises to explore concordance.

are asked to write down the different forms of words elicited by concordancing strings provided in the exercise; for example, the first chart in Figure 5 asks students to write down words such as *accurate*, *accurately*, and *accuracy* as they appear in concordances of the string "accur" in two different text files. In the second chart, students must provide their own search string and specify their own text files (this wouldn't have to be in one step; text file names could be removed one by one, for example). At this point, the students have been set free to explore the database on their own.

Having the power to theoretically explore the database is not quite enough, though. Most students will have realized by now that they can peruse the database for contextualized examples of difficult words, but this is only one possible mode of inquiry. Exploring the database for linguistic insight is another. However, it cannot be assumed that all students, by virtue of their engagement in language learning, will think like linguists and automatically

## Conduct your own research Complete the following charts with the missing information.

Text File	Different forms of accur
CHEM1989.TXT	
BIO-1988.TXT	
ŝ	
Text File	Different forms of

Figure 5. More exercises to explore concordance.

probe for provocative syntactic insights. Again they must be guided and, through example, shown that even mundane explorations can produce intriguing results.

At this point, the field is open for the materials developer, and the following is only a glimpse of the potential for inquiry available to students. But here are some of the things we do:

- 1. Students are asked to run a concordance on the word "much." They find that the words *how*, *very*, *and as* precede the word *much* in 75% of the items in one particular database. Students can be asked to look for exceptions, as when the phrase following *much* is adjectival rather than a noun phrase.
- 2. Students concordance the word *little* on a spoken English database and are asked to find the often repeated expression which means "a small amount" (the expression is "a little bit"). They are also asked to find from the data an intensifier for that expression (*tiny*, as in a little tiny bit). Finally, they are asked to compare the number of occurrences of little in the written and spoken databases. A similar exercise can be done with the word lot, in which case examples of a lot of, a whole lot, an awful lot, and quite a lot might occur in the concordance output.
- 3. Students run a concordance on the word *since* and must determine in what percentage of the occurrences that word means *because* and in what percentage it sets off a time marker in conjunction with the perfect tense.
- 4. Students concordance the word *such*, and are led to detect the following pattern: if the expression is *such* a then the noun following is singular, whereas following *such* alone, the noun is always plural. If there are any occurrences of *no such*, this is worth pointing out.
- 5. One fruitful area for concordancing is the use of auxiliary verbs, such as *do* and *does*, and modals such as *would*, *must*. Students can be asked to predict how often such words will be used in interrogative and how often in negative sentences, then do the concordance as a check on their intuition. The result may surprise them and reinforce the idea that concordancing can dispel unsubstantiated notions with hard data, but this is really only a device to get them looking closely at the output. There are many other tasks that can be built around such words.

### Conclusion

These few examples by no means exhaust what can be done with concordance data. It is interesting to speculate about what effects such exercises might have on student acquisition of grammar. As McDonough points out (1986; p. 31) regarding rules of grammar, "an imposed classification is usually less easy to remember and therefore less efficient than one invented for oneself." In theory, concordances should allow students to generate and collate the language data needed to invent their own rules of grammar—or to blaze their own neural pathways. Rumelhart and McClelland (1986; p. 217) discuss

the ways in which language learners postulate such rules. But what is clear is that even students with little computer experience and/or with discerning patterns in raw data can, with proper guidance, be successfully introduced to concordancing and thus started on the road to using concordance tools to explore a target language for purposes of their own.

(Received March 1990)

### REFERENCES

- Higgins, J., & Johns, T. (1984). Computers in language learning. Reading, MA: Addison-Wesley.
- Goethals, M. (1987). Automated analysis of the vocabulary of English texts and the generation of practice materials. Paper presented at the meeting of the 8th World Congress of Applied Linguistics, University of Sydney, 16–21 August, 1987, Sydney, Australia.
- Griffiths, R. (1990). An MS-DOS machine-readable corpus of transcribed scientific lectures. Available through the MS-DOS Users' Group, CALL-IS, TESOL.
- Johns, T. (1989a). Whence and whither classroom concordancing? In Bongaerts *et al.* (Eds.), *Computer applications in language learning* (pp. 9–33). The Netherlands: Foris Publications.
- Johns, T. (1989b). Should you be persuaded? Two samples of data-driven learning materials. Unpublished manuscript.
- McDonough, S. (1986). Psychology in foreign language teaching, second edition. London: Unwin Hyman.
- Rumelhart, D., & McClelland, J., (1986). On learning the past tenses of English verbs. In Rumelhart, D., J. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructures of cognition*. Vol. II, Psychological and biological models (pp. 16–271). Cambridge, MA: MIT Press.
- Stevens, V. (1988a). Studying vocabulary using concordances on microcomputers. TESOL Newsletter 22 (3),27.
- Stevens, V. (1988b). Misconception and misuse in CALL. CALL Digest 4 (4), 2-3

Vance Stevens, Instructional Developer, Sultan Qaboos University Language Centre, Oman, is Contributing Editor for *CALL Digest*, past-chair of CALL-IS in TESOL, Reviews Editor for *CALL-IS Newsletter*, and Chair and Newsletter Editor for the CALL-IS/MS-DOS Users' Group. He regularly publishes CALL articles, software, and bibliographies.