### Using Corpora to Teach and Assess Vocabulary

**Norbert Schmitt**

**Centre for Research in Applied Linguistics**

**University of Nottingham**

*Abstract

Corpus linguistics has been instrumental in redefining our understanding of how languages behave.  Many of these insights have direct pedagogical implications. This chapter illustrates how corpus evidence can be used by teachers to enhance their understanding of English usage and by students to learn English through inductive exercises.  Corpus analysis techniques can also be productively employed in vocabulary assessment.  Learner output can be analyzed for lexical diversity, lexical sophistication, and the use of certain types of word, e.g. academic vocabulary.  Future vocabulary test formats may even be able to measure collocational knowledge, based on the patterns identified from corpus research.*

Corpora have been a valuable resource in Applied Linguistic research in the last three decades, primarily providing insights into the frequency of occurrence of various language elements and the patterns of their use. Recently, several scholars have suggested that corpora may well be a valuable addition to second-language teaching methodology as well (Barnbrook, 1996; Carter and McCarthy, 1997; Johns, 1994; Reppon and Simpson, in press; Simpson and Swales, 2001; Wichmann, Fligelstone, McEnery, and Knowles, 1997).

CORPORA AND VOCABULARY TEACHING

Two obvious ways in which corpora can be used pedagogically are connected to deductive and inductive teaching approaches. In a deductive teaching approach, corpus evidence can be used to better inform teachers about the language elements they are presenting, and to provide clearer and more authentic examples of those elements. For example, Reppon and Simpson (in press: 100-111) show how concordance lines can help teachers understand the quite subtle differences between the forms *think of* and *think about*:

 **Imagine that you have been asked to explain the difference in use between *think of* and *think about***

•First, try to decide if through experience and intuition you can come up with a pattern for when one form is preferred over the other.

• Next,Now, look at the concordance lines provided below for *think of* and *think about*, taken from a corpus of informal spoken conversation. The target expressions, *think of* and *think about*,have been bolded in the concordance lines presented below. Pay special attention to what comes before and after the target words (i.e., *think of/about* what?). Are there any generalizations that can be made that would help a learner know when to use *think of* and when to use *think about*? To help you, the target expressions, *think of* and *think about*,have been bolded in the concordance lines presented below. To check your answers please go to the end of the Reference Section to see a summary of some of the generalizations about the uses of *think of* and *think about.*

 **THINK OF**

 stank. Then, as he was trying to **think of** something to say to her (all

 yes, wedding presents. We must **think of** something. You probably don't

racking my brains for three hours to **think of** something, I simply cannot last

 a second catastrophe. I tried to **think of** something to say myself, but my

 offered frills. Nicandra tried to **think of** something pleasing to say:

 only you were here, then we could **think of** something to do. "Christopher

 groaning quietly, perhaps trying to **think of** something that summed up what

 let said nothing. He had tried to **think of** something to say, but the only

 lunch? " " Ah me, the young! You **think of** nothing but your stomachs.

 sympathy and collusion. But I can **think of** nothing to say. Perdie says,

 she tried to speak, but she could **think of** nothing, and her mother, shifting

 anything so familiar, and he could **think of** nothing on earth to say. It

 man in the world. '" As he could **think of** nothing else, Martin repeated

 But try as she might, she could **think of** nothing to say like that, fierce

 listening. `Can we ourselves **think of** nothing that needs to be done?

 " what an idiot I was not to **think of** it before! You all right Elfie?

 .. no, wait a minute, come to **think of** it you 're finding. hmm.

 or him, on other occasions, come to **think of** it. We've been aware of each

happened to those kids. And come to **think of** it, Hamelin's rats and children

 like that five years ago, come to **think of** it, or even ten. It 's the

 wash his feet, he had seen, come to **think of** it, the moon not too remote from

probably cheaper than Selina, come to **think of** it, what with the hotel mark

 could have. I didn't happen to **think of** it then. 'And when did you

 her pregnant. Better not even to **think of** it. Just go on hating him,

 and done with. Don't let us ever **think of** it again. My family always

 "How nice. What did you **think of** it? " Patrice held her breath,

 **THINK ABOUT**

 <u idnum=607> You wouldn't just **think about** it it 's just gone isn't it

 Well that 's a good way, if you **think about** it he's got, he's got four

 more, I mean they can wear, if you **think about** it they were suits in the

 <u id=15 num=206> When you **think about** it, yeah he was So what '

 it seems easier that way when you **think about** it dunnit? Mm it's a lot be

does that come from? Oh when you **think about** it Pledge, why do they call

 wasn't the money really when you **think about** it because at end of day,

 more. I mean they can wear if you **think about** it they wear suits in the

 week! And why, they don't need to **think about**  it, they can talk you out of

 penetrating as lasers. `We might **think about** that, ' I say at last.

 I'll have to start and **think about** that train, Dwight.

 see it. That's the way I like to **think about** that sort of place. It's

 another way, but I don't want to **think about** that for a while. 'Timothy

 get eight to twenty - five. Now **think about** that. The district attorney XXX has,

Once teachers have examined concordance lines like this, they should be able to explain the differences in usage much more clearly and confidently. In this case, Reppon and Simpson suggest that *think of* is often used with indefinite references (e.g., *something*, *nothing*, and *it* referring to nothing in particular), while *think about* usually refers to more specific things (*it* and *that* referring back to specific references in the previous text). (See Reppon and Simpson, in press, page 293-294 for the complete analysis.)

Similarly, in an inductive approach, concordance lines from corpora can be used to provide the linguistic data from which learners can induce language rules and regularities for themselves. In this case it is important for the teacher to determine the level of the students and to select concordance lines that are both within the students' ability and which clearly illustrate the linguistic point(s) in question. Consider the following lines I adapted from the 2-million word British National Corpus Sampler using the WordSmith 3.00 concordance package:

 who had held the position **since** 1510

 the first non-Communist leader **since** 1948

 trying for their first win their **since** 1975

 he hasn't been back to work **since** Christmas

 this is their best plan **since** early February

 has not been seen **since** Friday

 I've been here **since** Saturday

 held the person **for** 10 hours

 is capable of lasting **for** 35 years

 did not score **for** 20 minutes

 leave it alone **for** a bit

 to comfort her **for** a brief moment

 has been going on **for** a century

 she stopped, but only **for** a moment

 the last dance went on **for** a long time

The above lines should make it relatively transparent for stronger beginning or lower intermediate learners that *since* involves a 'point in time' and *for* involves a 'duration of time'. Note that teachers may want to simplify a few of the words in the concordance lines, such as *communist* and *capable*. Beyond that, students should be able to evaluate these lines for themselves, once they have had some practice doing this type of analysis. The main advantage of this type of inductive exercise is that students can become linguistic 'Sherlock Holmes' and begin to look at the systematicity of language as an interesting linguistic puzzle, rather than a set of boring rules to be memorized. For many students, this approach can be more motivating and interesting.

Although they are not a magical solution to all problems (see Cook, 1998), there can be little argument that corpora used as above can add to second-language teaching methodology in a beneficial way. But in order to best refine effective teaching, we must also have effective assessment techniques in place to describe our learners' progress (or lack thereof). However, what should we assess? When it comes to vocabulary, all teachers are aware that their learners must know more than a word's meaning in order to use it well; they must also be able to use it appropriately. This entails knowing something about its stylistic constraints, and well as the way that it patterns with other words (collocation) (see Nation, 2001). Corpus evidence is the prime source for this information; as such, it is a key requisite for successful assessment of vocabulary use in context.

CORPORA AND VOCABULARY ASSESSMENT

Tests of receptive vocabulary knowledge are ubiquitous, because they have the key advantage of allowing the selection of the target words to be measured. They range from common multiple-choice formats, through matching formats such as the *Vocabulary Levels Test* (Schmitt, Schmitt, and Clapham, 2001), to self-report formats such as checklist tests (Meara and Buxton, 1987). Corpus evidence can provide information for the development of these receptive tests, including 1) the relative frequency of the target words, 2) the most frequent meaning sense of polysemous words, as well as 3) providing authentic examples which can be incorporated into the tests.

Productive vocabulary tests have been more problematic, mainly because it is difficult to combine authentic use contexts with the elicitation of preselected words. If we force examinees to use certain words, it is seldom in ways that are similar to real-world use. Consider two productive formats:

## **Give the L1 translation to these English words**:

1. sincere -

2. mandatory -

**Use the following words in a story:**

mandatory, sincere,

Neither of these tasks are particularly authentic. One would not normally give L1 translations while communicating in English outside the classroom; the whole idea is to avoid using translations while engaging in ESL discourse. Similarly, it would be very strange indeed to be required to use certain words when telling a story in the real world; if someone is relating a story, he has the freedom to tell it in his own way and in his own words.

If we take a more open approach to productive vocabulary testing, we need to analyze learner discourse where the task does not have predetermined lexical constraints. This approach entails collecting output from a learner and then analyzing it. The output can either be a single instance, or better, a combination of numerous and varied productions by the learner. Essentially the assessor is building a corpus of a particular learner's language output. This approach has high situational validity, in that students' language can be gathered while they are engaged in authentic tasks, such as making a list of items to take on an international trip or writing an academic paper for an actual class assignment. This approach has serious limitations when it comes to assessing lexical patterning (see below), but can work well for the measurement of various other lexical attributes, including the range of vocabulary used, the sophistication of vocabulary used, and the use of appropriate academic vocabulary.

## **Range of Vocabulary Used**

Let us use an example to illustrate these kinds of analysis, using an extract of my own writing on corpora from one of my books (Schmitt 2000: 69):

*It was when texts could be quickly scanned into computers that technology finally revolutionized the field. With the bottleneck of manually typing and entering texts eliminated, the creation of immensely larger corpora was possible. We now have 'third-generation' (Moon, 1997) corpora which can contain hundreds of millions of words. Three important examples are the COBUILD Bank of English Corpus, the Cambridge International Corpus (CIC), and the British National Corpus (BNC). The Bank of English Corpus has over 300 million words, while the CIC and BNC each have over 100 million. These corpora are approaching the size where their sheer number of words allow them to be reasonably accurate representations of the English language in general. This is partly because their larger size means that more infrequent words are included.*

The first step involves scanning or keying in the output into an electronic format. Once this is accomplished, a concordancing or other program can be used to analyze the vocabulary. The range of vocabulary can be analyzed with the word list function of a standard concordancing program, which will create either an alphabetic or a frequency list of the vocabulary used by the learner. Using the WordSmith concordancer, the above extract yields the following word lists:

SUMMARY

Tokens 130 Types 84 Type/Token Ratio 64.62

ALPHABETICAL WORD LIST FREQUENCY WORD LIST

Word Frequency Percentage Word Frequency Percentage

accurate 1 0.77 the 10 7.69

allow 1 0.77 of 8 6.15

and 3 2.31 corpus 4 3.08

approaching 1 0.77 words 4 3.08

are 3 2.31 and 3 2.31

bank 2 1.54 are 3 2.31

be 2 1.53 corpora 3 2.31

because 1 0.77 English 3 2.31

BNC 2 1.54 bank 2 1.54

bottleneck 1 0.77 be 2 1.54

One of the indicators of the diversity of the vocabulary in a text is the type/token ratio. It is calculated as follows:

 number of separate words (types)

Type/token Ratio = x 100

 total number of words in the text (tokens)

 (Laufer and Nation, 1995)

If most of the words are repeated several times, then fewer different words (types) have been produced by the learner. On the other hand, if few words are repeated, then more types will be included in the text. To give some indication of how to interpret a type/token ratio, Ure (1971) found that spoken texts generally had ratios under 40. Written texts generally had ratios over 40, although they ranged from 36 to 57. The ratio for this extract (64.62) is relatively high, indicating that a relatively large number of different words were used. In contrast, a relatively low ratio might indicate that a student was over-relying on a limited number of word types. (See Read, 2000 for a more advanced discussion of type/token ratios, including their limitations.)

The word lists can give a more precise indication of the vocabulary being produced by a learner. The frequency list is particularly helpful in illustrating the patterns of learner usage, i.e. which words are being used multiple times. With this list, a teacher can examine a learner's output and determine whether she is over-using some words when other words may be more appropriate. Likewise, the alphabetical list forms a handy reference which a teacher can save for comparison with future learner output.

**Sophistication of Vocabulary Used**

The sophistication of the vocabulary used can be determined by a software program which analyses the vocabulary used according to its frequency of occurrence in general English. One such program is RANGE, which gives an indication of the vocabulary used, dividing it into several categories of frequency: the most frequent 1000 words, the most frequent 1000-2000 words, academic words according to the Academic Word List (Coxhead, 2000), and all other words not on these three lists. The analysis of the above extract with this program looks like this:

WORD LIST TOKENS/% TYPES/% FAMILIES

1st 1000 93/ 73.8 58/ 69.9 50

2nd 1000 4/ 3.2 4/ 4.8 4

Academic words 11/ 8.7 10/ 12.0 10

Not in the lists 18/ 14.3 11/ 13.3 ?????

Total 126 83 64

Types Found In The 1st 1000 List

TYPE RANGE FREQ

THE 1 10

OF 1 8

WORDS 1 4

AND 1 3

ARE 1 3

ENGLISH 1 3

BANK 1 2

BE 1 2

HAVE 1 2

LARGER 1 2

Types Found In The 2nd 1000 List

TYPE RANGE FREQ

IMMENSELY 1 1

INFREQUENT 1 1

INTERNATIONAL 1 1

QUICKLY 1 1

Types Found In The Academic Words List

TYPE RANGE FREQ

TEXTS 1 2

ACCURATE 1 1

APPROACHING 1 1

COMPUTERS 1 1

CREATION 1 1

ELIMINATED 1 1

FINALLY 1 1

MANUALLY 1 1

REVOLUTIONIZED 1 1

TECHNOLOGY 1 1

Types Not Found In Any List

TYPE RANGE FREQ

CORPUS 1 4

CORPORA 1 3

BNC 1 2

CIC 1 2

BOTTLENECK 1 1

BRITISH 1 1

CAMBRIDGE 1 1

COBUILD 1 1

SCANNED 1 1

SHEER 1 1

THIRD-GENERATION 1 1

\*Note: The RANGE column indicates the number of texts the word appears in. This column would be important if more than one text was analyzed.

RANGE has the advantage of indicating how frequent the words appearing in a text are in general English. Research has shown that, in general, learners acquire more frequent words (like *boy*, *way*, and *see*) before words of a lower frequency (like *stimulate*, *cabinet*, and *dazzling*). This means that beginning/intermediate second language learners are likely to know mostly high frequency words, and only a limited number of low frequency words (this may not be true for advanced learners). In addition, researchers have found that the language of even proficient speakers is predominantly made up of high frequency vocabulary, with the most frequent 2,000 or so word families accounting for 80 percent or more of any written text (Nation and Waring, 1997). For spoken discourse, 2,000 word families can make up more than 98 percent of a typical conversation (Schonell, *et al*., 1956).

Since high frequency vocabulary is used by both proficient and nonproficient English users, one way to gauge the sophistication of vocabulary use is to determine how much of the vocabulary produced is of relatively low frequency -- beyond the 2000 frequency level. According to Laufer (1994) this vocabulary can be considered non-basic, and therefore can be viewed as roughly sophisticated vocabulary. In my extract, the most frequent 2,000 words of English make up 77% of the tokens and 74.7% of types. This means it contains a relatively high percentage of 'sophisticated' vocabulary: 23% of the total tokens and over 25% of the types. Learners are likely to have far lower percentages, but in general, the higher their percentages, the greater the tendency to use the more precise lower frequency vocabulary found in the 'academic' and 'not found in any list' categories.

## **Use of Academic Vocabulary**

RANGE also gives an indication of the academic vocabulary used. One problem with learner writing is that general vocabulary is used instead of the more precise academic vocabulary. Typical writing by proficient academic writers contains about 7-10 percent academic vocabulary. In addition, use of academic vocabulary is an important factor in giving a text an 'academic tone'. RANGE can give an indication of the degree to which learners are using academic vocabulary. If the percentage is too low, it is likely that learners this is affecting both the content and tone of their academic writing in negative ways.

VOCABULARY ASSESSMENT IN THE FUTURE: MEASUREMENT OF PRODUCTIVE COLLOCATIONAL KNOWLEDGE

The 'collect and analyze' approach discussed above works reasonably well when the lexical unit being targeted is either a type or word family. However, we do not yet have software that can read a text and reliably identify and isolate the collocational patterning which exists in language. This still takes human analysis of corpus results. The lack of an 'automatic' collocation pattern identifier is unfortunate given the current emphasis on appropriate vocabulary usage. One of the factors which largely determines the appropriate use vocabulary is its collocational behavior. If a learner can use a word in its typical collocational patterns, he stands a good chance of using it appropriately.

Unfortunately, assessors run into the same problem of receptive vs. productive tests when designing measurements of collocational knowledge. Corpus evidence can be analyzed to determine a word's collocational patterning, and then learners can be tested for knowledge of those patterns. However it is impossible in practical terms to analyze a free composition for all possible collocational patterns which may appear. Unless assessors are able to stipulate collocations and patterns in advance, any assessment of this sort becomes an *ad hoc* exercise which is unlikely to be practical or reliable. In other words, the field must develop some way of specifying appropriate collocations in advance and using that list of collocations in a scoring procedure. I have taken the first steps along this line of reasoning in a recent study (Schmitt, 1998). I noticed that the list of collocations for certain words seem to fall into certain semantic fields. For example, the collocates for the word *massive* included the following:

 *attack, damage, destruction, died, explosion, injuries, launched, military, refugees*

*amount, billion, budget, companies, debts, deficient, development, dollar, economic, expansion, financial, investment*

# Cause, changes, increased, in flux, reduced, rises, turned

 (Schmitt, 1998: 35)

It seemed to me that these collocates fell into areas that could roughly be described as war, economics or finance, and change respectively. Using these categories, I elicited sentences on these topics with the following prompts:

 Say a sentence using *massive*

1. if you were talking about war.

2. if you were talking about finance or the economy.

3. if you were talking about statistics.

With the topic for the target sentences being constrained in this way, I gave points for sentences which included collocates from the corresponding list. The procedure was partially successful, and I think that it has potential, but a number of problems still need to be resolved. First, it is not easy to identify a clear and sufficient listing of potential collocates to use as a norm list. Second, it is not clear whether each sentence produced by a learner needs only one collocate, or whether one than one should be required. Third, most of the collocates occurred within a five-word span of the target word (+/- 5 words in either direction), but some occurred more than 10 words away. Thus it is not clear how 'wide' a span to use in the assessment. Fourth, the procedure worked fairly well for typical collocates, but far less well for the less common collocates. I am still not sure of the eventual viability of this procedure, but until software programmers find a way to tag lexical patterns automatically (and will this ever happen?), I feel it is a direction worth pursuing.

CONCLUSION

Corpus procedures can be a great help in the teaching and assessing of vocabulary. In teaching, corpus evidence can be used in beneficial ways for both deductive and inductive activities. Perhaps its greatest benefit is in illustrating authentic language use in transparent ways.

In vocabulary assessment, corpus evidence is key to the development of most vocabulary tests. Software which facilitates the analysis of types and word families is now readily available, but the real vocabulary assessment prize remains elusive: the assessment of vocabulary according to its stylistic and collocational appropriacy. Regardless, however vocabulary assessment develops, corpora and corpus analysis will surely remain vital to any progress made.

NOTE

The software programs mentioned in this article are available at the following places:

The RANGE and WORD vocabulary analysis programs by Paul Nation are available free of charge at <[http://www.vuw.ac.nz/lals/](http://www.vuw.ac.nz/lals/.)>.

A very useful concordancing package is WordSmith Tools, which is available from Oxford University Press at <<http://www1.oup.co.uk/elt/catalogue/Multimedia/WordSmithTools3.0/>>.

Another concordancing package is MonoConc Pro, which is available at <www.athel.com>.

A good initial corpus is the British National Corpus (BNC) Sampler, which includes a 1-million word written sample and a 1-million word spoken sample from the complete 100-million word BNC. Source: <http://info.oup.ac.uk/bnc>

REFERENCES

Barnbrook, G. (1996). *Language and computers: A practical to the computer analysis of*

*Language*. Edinburgh: Edinburgh.

Carter, R. & McCarthy, M. (1997). *Exploring spoken language*. Cambridge: Cambridge University Press.

Cook, G. (1998). The uses of reality: A reply to Ronald Carter. *English Language Teaching Journal, 52*: 57-63*.*

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly 34*: 213-238.

Johns, T. (1994). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In Odlin, T. (ed.) *Perspectives on pedagogical grammar*. Cambridge: Cambridge University Press.

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

Laufer, B. and Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics 16*: 307-322.

Meara, P. and Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing 4*: 142-154.

Nation, I.S.P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nation, P. and Waring, R. (1997). Vocabulary size, text coverage, and word lists. In Schmitt, N. and McCarthy, M. (eds.) *Vocabulary: Description, Acquisition, and Pedagogy*. Cambridge: Cambridge University Press.

Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Reppon, R. and Simpson, R. (in press). Corpus linguistics. In Schmitt, N. (ed.) *Introduction to Applied Linguistics*. London: Arnold.

Schonell, F., Meddleton, I., Shaw, B., Routh, M., Popham, D., Gill, G., Mackrell, G. and Stephens, C. (1956) *A Study of the Oral Vocabulary of Adults*. Brisbane and London: University of Queensland Press/University of London Press

Schmitt, N. (1998). Measuring collocational knowledge: ket issues and an experimental assessment procedure. *International Review of Applied Linguistics 119-120*: 27-47*.*

Schmitt, N. (2000). *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press.

Schmitt, N., Schmitt, D. and Clapham, C. (2001) Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing 18*: 55-88.

Simpson, R. & Swales, J. (eds.) (2001). *Corpus linguistics in North America:Selections from the 1999 Symposium*. Ann Arbor: University of Michigan Press.

Ure, J. (1971). Lexical density and register differentiation. In Perren, G. E. and Trim, J.L.M. (eds) *Applications of Linguistics: Selected papers of the Second International Congress of Applied Linguistics, Cambridge, 1969*. Cambridge: Cambridge University Press.

Wichmann, A., Fligelstone, S., McEnery, T., Knowles, G. (eds.) (1997). *Teaching and Language Corpora*. London: Longman.

Biodata: Norbert Schmitt is Reader of Applied Linguistics at the University of Nottingham and co-director of the Centre of Research in Applied Linguistics. He has recently published *An Introduction to Applied Linguistics* with Arnold Press and is currently researching the acquisition of formulaic vocabulary sequences.