C us Studies in L uage Education

Edited by
Melinda Tan

IELE Press

Corpus Studies in Language Education

Edited by

Melinda Tan

IELE Press

IELE Press
(Institute for English Language Education)

Assumption University
Ramkhamhaeng soi 24,
Huamark, Bangkapi
Bangkok 10240
Thailand

ISBN 974-615-076-6 (Thailand)

Printed in Thailand

Contents

PART I: THE MENTAL LEXICON

PART II: EFL LEARNER CHARACTERISTICS

PART III: CONSCIOUSNESS-RAISING

PART IV: TEACHING AND LEARNING RESOURCES

PART V: REVIEWS

About the contributors

Guy Aston teaches English Linguistics at the University of Bologna in Italy. He has recently edited a volume on the use of corpora in language learning, *Learning with Corpora* (Athelstan, 2001), and is co-author (with Lou Burnard) of *The BNC handbook* (Edinburgh University Press, 1998).

Eric Bodin has a master's degree in TESL/Applied Linguistics from Iowa State University. He has taught writing and linguistics courses at freshman and graduate levels, at Iowa State University. He has also lectured on the International Teaching Assistant programme and taught English classes in Kazakhstan. His research interests include parallel concordancing in the classroom.

Ramesh Krishnamurthy has degrees in European languages from Cambridge and in Indian languages from London. He has worked at Birmingham University on the COBUILD project since 1984, writing dictionaries, grammars and other publications, and developing corpora and software.

Linda H.F. Lin is a teacher with 15 years of teaching experience. She has taught English to Chinese learners and also taught Chinese to English learners in China, Australia and Hong Kong. She now works at the English Language Centre of the Hong Kong Polytechnic University.

Yuanwen Lu is currently doing her PhD at the Department of English Language and Literature, National University of Singapore. Her research is a corpus-based study of the lexical behaviour in Chinese learner English. She obtained her M.A. From Shanghai Jiao Tong University, P.R. China and has taught EFL for several years at universities in China.

Alan Maley is Dean of the IELE, Assumption University, Bangkok. From 1962-1988 he worked for the British Council in Yugoslavia, Ghana, Italy, France, PR China and India. He was Director-General of the Bell Educational Trust, Cambridge, 1988-93, and Senior Fellow, National University of Singapore, 1993-98. He is series editor for the Oxford Resource Books for Teachers, and author of over 30 books on language teaching.

Michael McCarthy is Professor of Applied Linguistics at the University of Nottingham, UK, and Adjunct Professor of Applied Linguistics at the Pennsylvania State University, USA. He is author of many books and articles on vocabulary and on spoken discourse. He has lectured on language and language teaching in 36 countries and has been actively involved in ELT for 37 years.

Vincent B Y Ooi is Senior Lecturer in the Department of English Language and Literature, National University of Singapore. His many research interests include corpus linguistics, lexicography, varieties of English and lexical semantics.

Norbert Schmitt is Reader of Applied Linguistics at the University of Nottingham and co-director of the Centre of Research in Applied Linguistics (CRAL) at the University of Nottingham. He has recently published *An Introduction to Applied Linguistics* with Arnold Press. He is interested in all aspects of second language vocabulary, and is currently researching the acquisition of formulaic sequences.

Melinda Tan is a lecturer at the IELE, Assumption University, Bangkok. She is the Managing Editor of *The English Teacher: An International Journal.* Her research interests include applications of corpus linguistics in the language classroom, cognitive semantics and critical discourse analysis.

Christopher Tribble is an Associate Lecturer in Applied Linguistics in King's College, University of London, where he is responsible for courses in Discourse and Genre Analysis and Teaching of English for Academic Purposes. He has taught in primary, secondary and tertiary education in the UK, and has worked as a lecturer, teacher trainer and project manager in China, Eastern Europe and South Asia. He is the author of *Writing* in the OUP teacher education series and has a long-term interest in professional communication, evaluation in education, and the use of computers in text analysis and language description.

Xiaoling Zhang is currently a lecturer at the University of Nottingham, UK. She received her first and master's degrees from Shanghai International Studies University, China, where she worked as a lecturer and Associate Professor for nine years She won a scholarship to do her PhD at the School of English Studies, Nottingham University in 1995. Her research interests include corpus linguistics, discourse analysis and second language acquisition.

Acknowledgements

Many people have helped in their different ways in the production of this book. First, I would like to thank all the contributors to *Corpus Studies in Language Education.* I would also like to thank Alan Maley for his help and support without which this book would not have been possible. Finally, I wish to thank Mario Saraceni for his invaluable help in the formatting of this book.

Introduction

Corpus linguistics in language description and language education

Investigating language structure and use

Recent studies in corpus linguistics have shown that intuitions about language use are not always the best way to understand the nature and structure of the language itself. For many years, traditionalists have sought to describe language based on intuitive perspectives and not from facts. This has led to many misleading notions about language, the main one being that language is divided into two aspects - form and meaning.- thus leading to grammar and vocabulary being taught independently from one another in the language classroom. The assumed division between form and lexis has been held as a misconception from work conducted by linguists such as Sinclair (1991a, 1996), Leech (1991), Stubbs (1995, 1996, 1998), Moon (1994, 1998), etc, who have all shown that both language aspects are inextricably linked, contrary to previous intuitions. Sinclair, especially, has established his own position regarding meaning and form, asserting that:

> "each meaning can be associated with a distinct formal patterning...There is **no** distinction between form and meaning...[The] meaning affects the structure and this is..the principle observation of corpus linguistics in the last decade..." (Sinclair, 1991a: 6-7)

Here, it is clear that any change in grammatical choice causes a change in the lexical choice, and vice versa, consequently affecting the unit of meaning. This assertion about meaning is provocative because it claims that:

> "every sense or meaning of a word has its own grammar...each meaning can be associated with a distinct formal patterning..." (Sinclair, 1991a: 10)

The existence of a distinct formal patterning would imply the existence and inter-relation between the syntagmatic and the paradigmatic axes, which according to traditional perspectives, did not exist before. This inter-relation between the syntagmatic and paradigmatic axes is illustrated when the syntagmatic axis, which shows the combination of words -grammatically and simultaneously - *prospects* certain other words on the paradigmatic axis, whilst grammatically opening up certain classes of words on the paradigmatic axis (see Tognini-Bonelli, 1996).

Corpus-based approaches and teaching

The first question to be answered here is "What is a corpus-based approach?". Biber (1998) succinctly outlines the main features of a corpus-based approach. This approach follows from the characteristics which identify corpus-based analyses:

- it analyses actual patterns of language use in authentic data
- it makes use of a large collection of corpus data taken from written, spoken texts or both,
- it utilises computer concordancing and tagging programmes for the analysis.
- it relies on corpus linguistic principles of analysis to report findings.

Several of the advantages of corpus-based approaches can be found in teaching with respect to the following three aspects:

- Open-ended supply of language data
- Promoting discovery-based learning
- Customised teaching materials for learners

Open-ended supply of language data

Large and accessible supplies of language data are valuable as resource material which have been exploited to devise corpora as part of materials development in the delivery of computer-delivered learning packages. Besides the use of general-purpose corpora such as the BNC and COBUILD, corpus data have also been used in developing LSP (Language for Specific Purposes) corpora as well as other sublanguages such as computer manuals, applied sciences, language engineering, etc. Leech (1997) and Aston (1997) highlight the value of learning the linguistic characteristics about language varieties through these kinds of specific corpora, especially for people wanting to specialise in various fields. Thus lexical frequencies, collocations and characteristic grammatical structures are beneficial for a better understanding of a particular kind of language variety. For example, Aston (1997:52) focuses on the value of newspaper corpora for use in "selecting texts with particular characteristics and smaller contexts for illustrating a particular linguistic phenomenon".

The rationale behind employing corpus data for teaching purposes follows from the view that learners would be able to reproduce authentic language behaviour from naturally occurring texts (see Sinclair 1991a). The larger the amount of naturally occurring texts, the better the evidence for a more accurate description of the characteristic features of language. Sinclair reiterated this view in 1997 when he said:

> "In order to uncover the regularities of structure, to identify, if possible, exactly what the realisations are of meaningful choices and to give precise shape to all the linguistic categories of linguistic description, it is necessary to assemble a large number of putative instances of each phenomenon. Given the well-known distribution of word tokens in a language, a large corpus or collection of texts is essential to provide a body of evidence" (Sinclair 1997:28)

The implication of using large amounts of corpora for language teaching is that learners can use the evidence in corpora for introspection. Introspection is seen to be a behaviour desired from learners where learners are viewed as active participants of language from a textual and discoursal perspective.

The view corpus linguists hold that large supplies of corpora would help create the desired language behaviour of learners as active participants in language is constantly challenged. Opponents argue that large amounts of computer data ignore aspects of culture and pyschology. Data cannot replace the complex mental processes that occur in meaning interpretation, organisation and classification of language in a learner's mind. Thus, even with large amounts of corpora, corpus analysis can only give a partial description of language since corpus linguistics 'comes from the perpective of the observer looking on, not the introspective of the insider' (Widdowson

2000:6). Many opponents against the use of corpora in language teaching have also raised questions about using native-speaker models (e.g. southern British English or American English) as evidence of attested language use to be taught tο language learners, citing linguistic imperialism or a conspiracy to impose English globally as reasons. Cook (1998, 2001) and Widdowson (2000) highlight also the neglect of corpus linguists to consider appropriacy in contexts and providing choices to learners. They emphasise the need to give learners choices and opportunities to make their own impact in language as long as the expressions they use are appropriate in a particular context, even if the expressions have been learnt from invented sentences. This is more important than uttering memorised lexical phrases which are contextually inappropriate.

Criticisms about the use of corpus data in the classroom are usually based on a fear that corpus linguists are advocating corpus-based analysis as the only correct approach to language teaching. However, many of these criticisms are based on misconceptions about important precepts in corpus lingusitics (see Stubbs 2001). A corpus-based approach to language teaching is meant to be a complementary approach to traditional teaching approaches. What corpus linguistics offers for education are evidence, from data, of discoursal, socio-cultural and psycholinguistic insights which provide direct applications and even restructuring of syllabuses and materials for teaching and evaluation (see Higgins and Johns 1984; Johns 1988, 1991a, 1991b, 1993, 1997; Tribble and Jones 1990; Aston 1997; Carter and McCarthy 1997; Carter 1998a; McCarthy 1998; Schmitt 2000, 2001 and Tribble 2000). In defence of using native-speaker models such as English as evidence of attested language usage, Carter (1998a) is of the view that because most expressions especially in English are culture bound, there is a need to keep "cultural particulars" (Carter 1998a: 50) intact so as to promote an awareness of language in terms of sensitivity and cultural understanding.

Promoting discovery-based learning

Sinclair (1997) presents an important precept for language teachers based on corpus principles: *Present real examples only.* The rationale behind presenting only real examples comes observation that in the past coursebook writers have always relied on their intuitions rather than made use of authentic language data. Carter (1998a) explains, "the language of some coursebooks represents a 'can do' society, in which interaction is generally smooth and problem-free" (Carter 1998a: 47). Many corpus-based teaching and learning materials approaches have been designed to take into account authentic language use. These materials are clearly indications of the designers' reflections about the value of a corpus-based approach to pedagogy in which the focus is to match theory with practice and to "fashion pedagogic reality to fit the descriptive findings" (Aston 1995). Practically all of the corpus materials designed concentrate on an approach in which discovery-based learning is valued and language awareness with regard to sensitivity about language use is activated. As McCarthy (1998:23) suggests, "It is thus only when good observers of language combine their talents with the display and analysis of data by the computer that the optimum gains can be made." Of course, it is not enough just to leave the students to interpret the data and to expect them to analyse or introspect about language use, regularities of language, etc. Students need to be trained how to interpret and analyse the data presented.

Customised teaching materials

Corpus based teaching promotes the value of learning language in chunks rather than as single words based on Pawley and Syder's (1983) study about native-like selection and native-like fluency. Teaching the unit of meaning as being phrasal rather than the single word is at the core of corpus-based language tasks. One of the main advantages of designing customised teaching materials using corpus data is that data can be sequenced and graded to suit the linguistic level of the learner in the preparation of tasks involving the use of investigative skills. Depending on the pedagogical aim of the lesson, teachers can select and present as much corpus data as they require and quickly, from commercially available or on-line corpora (e.g. BNC, COBUILD) for use in the classroom. The data not only show patterns of real language use but are also based on real contextualised examples of written and spoken language.

Corpus-based teaching materials are usually designed with the following aims in mind. Language learners should be able to to do three things:

a) be consciously aware of the unfamiliar usages of language they have heard or read in native speaker contexts,

b) investigate how these unfamiliar usages are employed in natural authentic communication, and finally,

c) experiment with these usages in spoken or written communication, so that they become familiar.

Learners who have achieved these aims should develop sensitivity to language patterns of use in authentic communication. Learners are provided with authentic examples of unfamiliar English usages taken from real language interactions and through discovery learning, comparisons can be made with the learners' own language, cultural and world knowledge or experiences. Depending on the level of the learner, awareness of these differences can later be heightened to empower him or her to make informed choices of language use, according to the situation. (For further discussion, see Tsui, 1994; Carter, 1998a; McCarthy, 1998; McCarthy and Carter, 1994, 1995). In a sense, corpus-based approaches to learning teach the descriptive aspects of language and make the target language an object of study. Hence, knowing a word involves knowing its grammar as well and this involves knowing the patterns of use in which the word is regularly found. Corpus-based activities would aim to activate and assess such knowledge through the direct or indirect teaching of corpus linguistics principles such as collocation and colligation (see Hoey, 2000; Lewis, 2000; Schmitt, 2001). These principles are meant to help language learners develop analytical skills aimed at providing a better understanding about the nature and structure of language, and their own mental lexicons.

This volume

The papers in this collection are held together by the link that they have with corpus studies in various aspects of language education. Although the contributors have analysed corpora of various sizes in their respective investigations, (BNC[1], BoE[2], CANCODE[3], CLEC[4], JPU[5] Corpus, LOCNESS[6]) it is hoped that the fascinating insights provided by their findings regarding the nature of language and learning will be enough justification for purists who contest the reliability of data from small corpora.

The papers have been divided into five main sections:

Part I : the mental lexicon,

Part II : EFL learner characteristics,

Part III: consciousness-raising

Part IV: teaching and learning resources.

Part V : reviews

In Part I, the two papers by Michael McCarthy and Norbert Schmitt examine how corpus analysis reveals insights into the mental lexicon, especially with regard to vocabulary learning and testing. McCarthy focuses specifically on advanced level vocabulary and the problems faced by language learners of English in processing this level of vocabulary. Schmitt looks at how corpus analysis techniques can be used to measure the diversity and richness of vocabulary knowledge in the language learner's mental lexicon.

In Part II, the two papers by Yuanwen Lu and Linda Lin analyse patterns of misuse and overuse through comparisons between learner and reference corpora. Lu explains her findings through corpus analysis of how the lack of 'native-like' quality in Chinese Learner writing is a result of the overuse of expressions which are directly translatable from Chinese. Lin's paper offers further insights into the language characteristics of Chinese Learners of English by examining the co-textual conditions in which the pronoun "it" is misused by these learners.

In Part III, Xiaoling Zhang's paper on echoing and Melinda Tan's paper on prepositional clusters, focus on the language description of particular linguistic phenomena in authentic communication which are taught ineffectively as a result of the use of inauthentic data in the EFL classroom. The two papers make use of corpus data to highlight their claims and emphasise the role of consciousness-raising in developing observant and sensitive language users. Both papers also offer pedagogical materials, based on corpus data, for teaching their respective linguistic phenomena.

In Part IV, the five papers focus on the applications of corpora in the design of various resource tools for language teaching. Ramesh Krishnamurthy explains how the COBUILD English Dictionary (3rd edition) can be used to teach various language aspects such as vocabulary, pronunciation, phonetics, spelling, etc., as well as for correcting errors. Guy Aston elaborates on how a corpus such as the BNC can be used as a complement to other resource tools like dictionaries. He suggests that foreign students should be encouraged to access a corpus of the target language so that they can learn further information about the language and culture as well as promote independent learning. Continuing this focus on using a corpus as a resource for language education, Chris Tribble illustrates how a newspaper corpus can be a valuable tool in cultural studies through a keywords approach. Vincent Ooi, on the other hand, concentrates on describing other types of resource tools associated with corpus analysis such as concordancers and taggers.

He explains how concordancers such as the WordSmith Tools programme and taggers such as CLAWS could be beneficial for literary and linguistic analysis as well as for motivating student learning. Eric Bodin elaborates further on the value of using concordancers in the classroom based on the teaching goals. He offers a checklist of suggestions on what teachers should bear in mind when they decide to use concordancers e.g. type and size of corpus, linguistic features to be studied and the type of presentation.

In Part V, the link between the two book reviews is that of authentic language communication. Alan Maley's book review of David Crystal's (2001) *Language and the Internet* assesses how the Internet could serve as a useful resource for the linguistic description of a particular variety of authentic language use called Netspeak. Melinda Tan's review of Michael Lewis' (2000) edited book on *Teaching Collocation: Further Developments in the Lexical Approach,* on the other hand, evaluates the extent to which the lexical approach is applicable in the teaching of authentic language use in the classroom through corpus analysis tools such as collocation and colligation.

This volume is the first of its kind in South-East Asia. It is intended as a move in an on-going exchange of ideas and suggestions about language description and pedagogy between language practitioners in the East and the West. Overall, the aim of this book is, hopefully, to inspire readers to develop their own research questions regarding the nature of language use and the learner's mental lexicon, and to conduct their own corpus-based investigations in order to gain a richer understanding of these aspects.

Notes

[1] BNC is an acronym for the British National corpus. Further information about the corpus can be found at http://sara.natcorp.ox.ac.uk/

[2] BoE is an acronym for the Bank of English Corpus. Further information about the corpus can be found at http://titania.cobuild.collins.co.uk/boe_info.html

[3] CANCODE is an acronym for the Cambridge and Nottingham Corpus of Discourse in English. CANCODE is used as part of the Cambridge International Corpus (CIC). Copyright resides with Cambridge University Press.

[4] CLEC stands for Chinese Learner English Corpus. Copyright resides with Shanghai Jiao Tong University

[5] Janus Pannonius University (JPU) Corpus is a collection of Hungarian university students' writing in English. Further information about the corpus can be found at http://www.geocities.com/writing_site/thesis

[6] LOCNESS is an acronym for Louvain Corpus ofNative English Essays which is a subcorpus of the International Corpus of Learner English (ICLE) set up by the Catholic University of Louvain, Belgium. Further information about the LOCNESS corpus can be found at: http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/locness1.htm

References

Aston, G. (1995). Corpora in language pedagogy: matching theory and practice. In G. Cook and B. Seidlhofer (eds), *Principles and Practice in Applied Linguistics*. Oxford: Oxford University Press.

Aston, G. (1997). Enriching the learning environment: corpora in ELT. In A. Wichmann, *et al.* (eds), *Teaching and Language Corpora.* London: Longman.

Biber, D., S. Conrad and R. Reppen (1998). *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Carter, R. (1998a). Orders of reality: CANCODE, communication and culture. *English Language Teaching Journal* 52 (1): 43-56.

Carter, R. and M: McCarthy (1997). *Exploring Spoken English.* Cambridge: Cambridge University Press.

Cook, G. (1998). The uses of reality: a reply to Ronald Carter. *English Language Teaching Journal* 52 (1): 57-63.

Cook, G. (2001). The philosopher pulled the lower jaw of the hen. Ludicrous invented sentences in language teaching. *Applied Linguistics* 22 (3): 366-387.

Higgins, J. and T. Johns (1984). *Computers in Language Learning.* London: Collins.

Hoey, M. (2000). A world beyond collocation: new perspectives on vocabulary teaching. In M. Lewis (ed.), *Teaching Collocation: Further Developments in the Lexical Approach.* Hove: LTP.

Johns, T. (1988). Whence and whither classroom concordancing? In T. Bongaerts, P. de Haan, S. Lobbe and H. Wekker (eds), *Computer Applications in Language Learning.* Dordrecht: Foris, 9-27.

Johns, T. (1991a). Should you be persuaded - two samples of data-driven learning materials. In T. Johns and P. King (eds), *Classroom Concordancing (Special Edition), English Language Research Journal* 4: 1-16.

Johns. T. (1991b). From printout to handout: grammar and vocabulary teaching in the context of data-driven learning. In T. Johns and P. King (eds), *Classroom Concordancing (Special Edition), English Language Research Journal* 4: 27-45.

Johns, T. (1993). Data-driven learning: an update. *TELL and CALL* 3.

Johns, T. (1997). Contexts: the background, development and trialling of a concordance-based CALL program. In A.Wichmann *et al.* (eds), *Teaching and Language Corpora.* London: Longman.

Leech, G. (1991). Corpora and theories of linguistic performance. In Svartvik, J. (ed.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 92, Stockholm, 4-8 August 1991.* Berlin: Mouton de Gruyter, 105-122.

Leech, G. (1997). Teaching and language corpora: a convergence. In A.Wichmann *et al.* (eds), *Teaching and Language Corpora.* London: Longman.

Lewis, M. (ed.) (2000). Learning in the lexical approach. In M. Lewis (ed.), *Teaching Collocation: Further Developments in the Lexical Approach.* Hove: LTP.

McCarthy, M. (1998). *Spoken Language and Applied Linguistics.* Cambridge: Cambridge University Press.

McCarthy, M. and R. Carter (1994). *Language as Discourse: Perspectives for Language Teaching.* Harlow: Longman.

McCarthy, M. and R. Carter (1995). Spoken grammar: what is it and how do we teach it? *English Language Teaching Journal* 49 (3): 207-218.

Moon, R. (1994). The analysis of fixed expressions in text. In M. Coulthard (ed.), *Advances in Written Text Analysis.* London: Routledge, 117-135.

Moon, R. (1998). *Fixed Expressions and Idioms in English.* Oxford: Oxford University Press.

Pawley, A. and F. H. Syder (1983). Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In J. C. Richards and R. W. Schmidt (eds), *Language and Communication.* Harlow: Longman.

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

Schmitt, N. (2001). Using a word knowledge framework to analyze vocabulary tests and activities. In F. Fernández (ed.), *Los Estudios Ingleses. Studies in English Language and Literature* 3.

Sinclair, J. (1991a). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, J. (1996). The search for units of meaning. *Textus* IX: 75-106.

Sinclair, J. (1997). Corpus evidence in language description in A. Wichmann *et al.* (eds), *Teaching and Language Corpora.* London: Longman.

Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language* 2 (1): 23-55.

Stubbs, M. (1996). *Text and Corpus Analysis: Computer Assisted Studies of Language and Culture.* Oxford: Blackwell Publishers.

Stubbs, M. (1998). A note on phraseological tendencies in the core vocabulary of English. *Studia Anglica Posnaniensia* 23: 399-410.

Stubbs, M. (2001). Texts, corpora, and problems of interpretation: a response to Widdowson. *Applied Linguistics* 22 (2): 149-172.

Tribble, C. (2000). Practical uses for language corpora in ELT. In P. Brett and G. Motteram (eds), *A Special Interest in Computers: Learning and Teaching with Information and Communications Technologies,* Kent: IATEFL, 31-41

Tribble, C. and G. Jones (1998). *Concordances in the Classroom.* Houston: Athelstan.

What is an advanced level vocabulary?

Michael McCarthy
University ofNottingham

Abstract

This paper is concerned with the nature of advanced-level vocabulary, both in terms of description and second-language pedagogy. Using the 5-million-word CANCODE spoken corpus and a 5-million-word written sample of the Cambridge International Corpus, the paper isolates a band of low-frequency vocabulary occupying positions 6,000 to 10,000 in the frequency rank list of the combined corpora, and examines the types of words found and what problems they raise for learners of English. At this level, cultural aspects of meaning, collocability and more diffuse semantic prosodies come into play. There are also some problems concerning a lack of fit between frequency and psychological organisation of the lexicon which need to be addressed in designing an advanced level vocabulary syllabus.

Introduction

In a previous paper (McCarthy 1999), I argued for the use of large-scale corpus-based investigations to determine the size and nature of a basic level vocabulary for everyday spoken English. Using the 5-million word CANCODE corpus of spoken English from the islands of Britain and Ireland, I attempted to demonstrate that a core vocabulary of a little less than 2,000 words seemed to be working harder than all the other words, occurring as they do with much greater frequency than all the others and encoding some of the most fundamental types of meaning and pragmatic function in everyday face-to-face communication[1]. The implications of that paper were that the remaining, non-core vocabulary occurred relatively infrequently but was massive in size (anything between 30-50,000 word-forms being in circulation in everyday talk, and considerably more in typical written texts, perhaps 80,000). For language pedagogy, this long tail of vocabulary is something which cannot be simply left to an unorganised free-for-all. Even when the core vocabulary has been consolidated, and added to by, maybe, another couple of thousand words in post-elementary and intermediate teaching over a period of two to three years, there remains a huge number of words to be learnt, certainly too many to teach and practise in any secondary or tertiary English-language programme. This raises dilemmas for pedagogy, some of which will be addressed in this paper.

Most language teachers will, at some time or other, be faced with the problem of what, and how to teach at the advanced level. The questions uppermost in their minds are likely to be:

- How many words should my learners be able to understand and/or use?

- Given the impossibility of teaching all the low frequency vocabulary of a language like English, which words should be focused on?

- What types of knowledge should learners be developing at this level?

- How can language pedagogy assist learners to become independent and autonomous so that they can continue with the daunting task after they have left the security of the classroom and their organised language-learning environment?

The present paper will attempt to provide at least some answers and guidelines in response to these challenges.

Setting the parameters: what are the targets?

One of the things it is possible to do to either manually (albeit time-consuming but indeed carried out in the past by dedicated researchers prior to the advent of computers) or automatically is to assess how many words one needs to know (at least passively/receptively) to understand a given percentage or proportion of the words in any typical, everyday, but randomly chosen text. For example, one might have as a pedagogical target that a new text should be understood 90% by a group of learners without the intervention of coursebooks, dictionaries, glossaries or direct teaching by the teacher. In other words, the new learning burden should not be more than 10 percent of the lexical content of the text. Research shows that, in relation to this target, a working vocabulary of somewhere in the region of 6,000 words will ensure around 90% comprehension of typical (non-specialist) written texts in English (see Carroll, Davies and Richman, 1971). A vocabulary of 6,000 words means adding another 4,000 to the core, basic 2,000 words. This is by no means beyond the reach of typical groups of learners, and a levels-based programme of vocabulary-teaching materials such as the Cambridge *Vocabulary in Use* series works on just such a set of targets, with increments of approximately 2,000 words being offered at each of the levels from Elementary to Lower Intermediate to Upper Intermediate, targets based on a combination of corpus-based quantitative research and large amounts of feedback from teachers, learners, reviewers and pilot editions (see McCarthy and O'Dell, 1999 and 2001).

90% coverage of most texts sounds like good news for all, except that, as I have already suggested, a considerable pedagogical burden will be generated by the remaining 10 per cent of the lexical material (a) because it will be of relatively low frequency, and (b) because it will carry a lot of specific content meaning, given that about 80% of the text will be covered by the first 2,000 core words, which are relatively general in lexical meaning, or else delexicalised (verbs such as *get, do*; nouns such as *thing, person*) or else functional (grammar words, discourse markers, etc). Earlier discussions of the problem of low frequency vocabulary and text processing have recognised this dilemma (see Richards 1974; Honeyfield 1977). What is more, simply 'teaching' the missing ten per cents of new texts will not necessarily encourage the autonomous learning skills that will be necessary when learners go out into the world and continue to meet unknown words. A working, receptive vocabulary of 6,000 words, then, would seem to be a good threshold at which to consider our learners ready to embark on an advanced level programme which will have several aims:

- To push the vocabulary size towards comprehension targets above 90% (e.g. to 95%) for typical texts.

- To expose the learners to a vocabulary that corresponds to frequency levels beyond the first 6,000-word band, but which is not too obscure to be tantamount to useless.

- To impart the kinds of knowledge essential for using words at this level of sophistication, given their specific lexical content.

- To develop awareness and skills that will stand the learner in good stead for becoming an autonomous vocabulary-learner.

Vocabulary size

Pushing the vocabulary size to a level where 95% comprehension can be achieved unfortunately does not just mean adding another 2,000 words to the 6,000 ideally possessed by upper intermediate learners, since the frequency curve falls off dramatically to a point where almost everything is very low frequency indeed, even in massive corpora. The crude truth is that, the nearer one attempts to approach native-speaker levels of competence, the bigger the gap to leap. Figure 1 shows the increments needed to go from 90% comprehension to 95% and to 97% (highly-advanced, expert-user level). The leaps required are not evenly spaced.

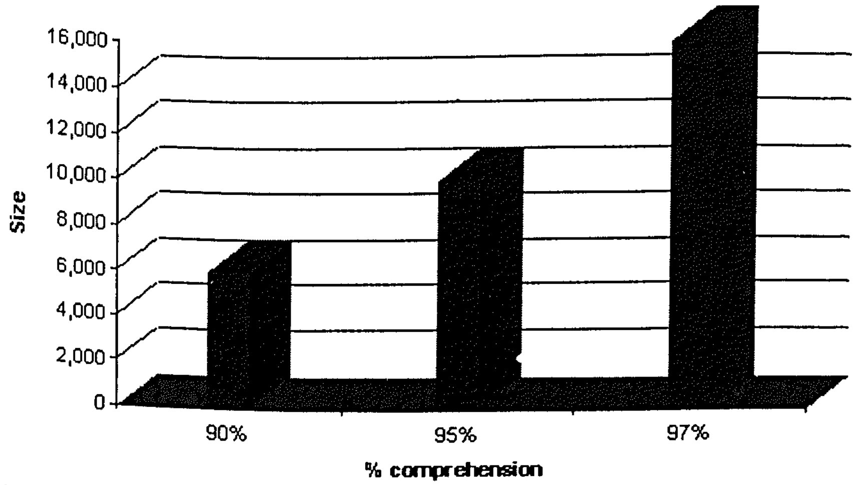


Figure 1: Vocabulary size and percentage of text comprehension

4,000 words (from the 6,000 to 10,000 word level) account for a 5% gain in comprehension, but the next 6,000-word increment (from the 10,000 to 16,000 word level) only brings with it another 2% comprehension gain, and so on. These figures are conservative and are taken, at this level, as excluding basic function words (non-lexical words). Depending on the type of texts and their level of specialisation, totals of 1,000 words either side of these figures may not be unexpected.

Our optimism about pedagogy at the 90% text coverage level (i.e. a working vocabulary of around 6,000 words) must be tempered by the stark reality that every tenth word in a typical new text will be unknown to our learners, and this is likely to be highly demotivating: there will simply not be enough known matter to support the guessing, inferring and deducing of meaning of the new matter. No one can reasonably be expected to look up one word in every ten in a dictionary and still remain motivated at the end of reading a 500-word text (50 look-ups). Research in press by Hu and Nation (reported in Nation, 2001: 147) supports the argument that a 90% text comprehension level is insufficient for a learner-reader to gain access to the text's message. Nation further argues (*ibid*: 147-8) that for true, pleasurable engagement with the meaning of a text, comprehension in the region of 98% is the threshold, clearly something that the average 6,000-word-level learner can only achieve with considerably simplified or very carefully selected material. The 95% comprehension level (suggested as being a vocabulary size in excess of 10,000 words by Figure 1, above) brings learners much nearer to a meaningful engagement with the content of a new text: one in 20 words will be new, but the contextual support and the motivation to look up new words will be massively greater. Carver (1994) suggests that native users operate at a 99% level of comprehension with average reading materials; clearly our learners cannot quickly or easily achieve that kind of level, but the 95% level (10,000 word-vocabulary) is, I would argue, achievable at tertiary level with good extensive reading programmes and intensive vocabulary teaching materials designed to focus on word knowledge at the 6-10,000 word-band level. The 6,000 to 10,000 word level, then, seems to be a key area where the gain in comprehension is still considerable; we have not yet reached the vast plain of rare vocabulary that adds little to overall comprehension potential. The 6- to 10,000 word level thus constitutes the main arena for the development of the advanced level vocabulary posed in the title of this paper.

The vocabulary of the 6,000 to 10,000 word band

The present paper uses the 5-million word CANCODE spoken corpus as one of its data sources. CANCODE stands for 'Cambridge and Nottingham Corpus of Discourse in English'. The corpus was developed at the University of Nottingham, UK, and funded by Cambridge University Press, with whom sole copyright resides. The corpus conversations were recorded in a wide variety of mostly informal settings across the islands of Britain and Ireland, then transcribed and stored in computer-readable form. Details of the corpus and its design may be found in McCarthy (1998). In addition, for this study I used the larger resources of the Cambridge International Corpus (CIC) (also copyright Cambridge University Press) to consult a 5-million word written corpus of mixed texts taken from newspapers, magazines, popular fiction, letters, etc., which constitute a broad-based everyday written corpus alongside CANCODE.

It is a straightforward matter to isolate the 4,000 words which occur between frequency ranks 6,000 and 10,000 in the overall corpus frequency list. That list cannot be presented in its entirety here for copyright reasons, but its content and flavour will be the subject of broad description and discussion in this section.

Figure 2 shows how these words are distributed in terms of gross frequency of occurrence in the 5-million word written corpus. It can be seen, for example, that 461 of the words occur more than 60 times in our sample, but that over 3,000 of the 4,000 are only occurring 40 times or less. Nonetheless, the frequency curve is relatively stable, with even the words in the 8-10,000 word rank occurring with sufficient frequency not to consider them rare or useless. 30 or 40 occurrences of a word are usually sufficient for robust patterns of usage and meaning to emerge in corpus concordances. It should be noted, though, that in the same corpus, even the bottom 100 of the core top 2,000 word list are occurring more than 250 times, so the frequency rates are very relative.

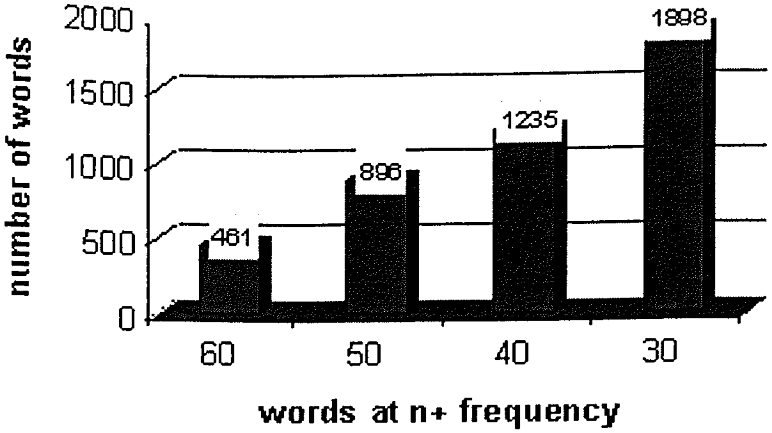


Figure 2: Frequency distribution 6000-10000 band (5 million written)

The vocabulary in this band is, unsurprisingly, rather varied. In the semantic field of clothing, for example, we find *stockings* and *tights*, useful items to be added to the already learnt basic clothing items such as *jacket* (word rank 1286) and *trousers* (word rank 2286). But we also find *shoe*, an apparent anomaly given the high rank of its plural *shoes* (word rank 1640). Such puzzles are not without linguistic interest. In everyday talk and writing, most of our references are indeed to *shoes* in the plural; the singular *shoe* might well be confined to more specialised contexts (its 65 occurrence in the written corpus are mainly fictional/literary ones). Quite clearly sensible pedagogical decisions come into play here, and there will be no need to introduce the singular form shoe as a 'new item' since its meaning can be deduced from one's knowledge of the prototypical plural form. Equally, corpus findings must always be tempered by the psychological notion of associated sets. Learners can and do learn sets of words which are semantically or psychologically associated, regardless of their difference in frequency of occurrence. For this reason, many learners may have already acquired *tights* and *stockings* in a learning activity at a much lower level based on, for example, an illustration of a person with all his/her clothes labelled with their English names. Perhaps a more obvious case is the word *Belgium*, which is of low frequency and occurs in the 6,000-10,000 band. One might speculate on whether British and Irish speakers hardly ever talk about Belgium, but few would deny the likelihood and usefulness, in a European context, of at least teaching the names of the countries of the European Union all together, and not relegating Belgium to the level of an 'advanced' word.

Frequency of form is misleading in another sense too. Although *spine* occurs in the 6,000-10,000 word list, not all of its meanings are as 'part of the human body', and metaphorically extended meanings such as 'part of a book where the binding is attached' or 'main vertical item in a network' (as in 'spine of a national network of cycle routes) occur. In this case, *spine* may well have been learnt as a body part at a lower level and as part of a psychological set, but we may indeed wish to revisit it at the advanced level in its extended meanings. Indeed, much advanced level vocabulary teaching will be a question of dealing with extended and metaphorical meanings, and new psychological sets may be forged which are once more at odds with frequency. For example, *spine* forms part of a set with *jacket* as belonging to the field of 'books', even though the meaning of *jacket* as a book-part does not occur in the present corpus. New associations may need to be forged, as in Table 1:

existing learner set	new learner set	existing learner set
spine head back thigh neck etc.	**spine** **jacket** binding cover frontispiece etc.	**jacket** trousers shirt skirt sweater etc.

Table 1

A further case of the mismatch between frequency and psychologically associated networks is the fact that some psychological sets in the same frequency band seem to have gaps. For example, our 6,000-10,000 band has *clasp, cling, clutch, grip, grasp*, but not *grab* (which has a much higher rank, occurring more than 300 times, putting it into a lower learning level band). The teacher may need to ensure either that the learners already know *grab*, or else import it to the advanced level as a 'new' word to make the set coherent. Materials designers would probably wish to include it in the set anyway, as part of a psychologically satisfying whole.

Another important aspect of frequency at this level is the occurrence of fixed expressions. The top 2,000 core words combine to form everyday fixed expressions of high frequency (especially in the CANCODE spoken corpus), such that frequencies of the following items are enough to rank them way above most of the single word-forms in our 6,000-10,000 word list, as in the sample in the table below:

Expression	**Frequency in 10m words (CANCODE + written)**
as I say	585
get rid of	471
have a word with	110
when it comes to	106

At the advanced level, fixed expressions continue to emerge, but are now more likely to be the semantically opaque, idiomatic ones. Their occurrences are likely to be low, but their meanings challenging, and their presence in texts highly psychologically salient. This is one of the paradoxes of lexical learning: rarity often increases salience. The phrasal verb *show up*, with its several idiomatic meanings, occurs more than 100 times, and the idiomatic phrase *on the spot* occurs 37 times in our sample, bringing both into the same frequency levels as the single-word 6,000-10,000 word list. Realistically too, because they are inherently less frequent, pedagogy will need to broaden its scope at this level and to make a wider trawl of the frequency list to include idioms of lower than 30 occurrences. *Peace and quiet*, for instance, occurs 24 times, and is typical of many binomial structures with occurrences of between 10 and 30 in our sample (see the concordance below). Account will have to be taken, too, of widely divergent frequencies in speech and writing, and for idioms in this range of frequency the corpus under investigation may need to be considerably enlarged before a clear picture can be seen to emerge. For example, the two idiomatic expressions *stumbling block* and *it just goes to show* have widely differing frequency patterns in speech and writing, but more than our total 10 million CANCODE plus written sample is needed to demonstrate this adequately. Figure 3 is based on the addition of the 10-million word spoken element of the British National Corpus to the CIC and CANCODE samples (figures are occurrences per 10 million words):

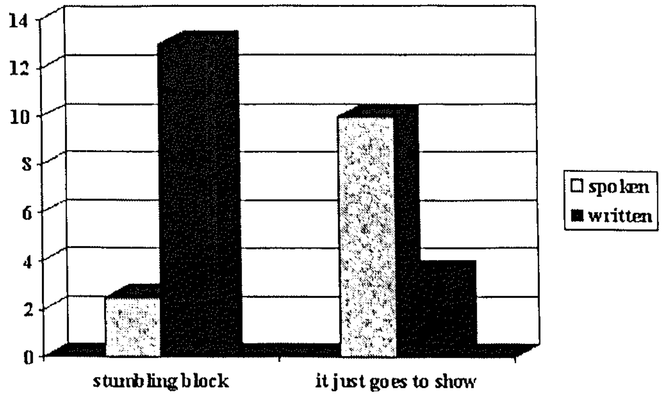


Figure 3

Our overall conclusion regarding the vocabulary of the advanced level frequency bands must be that the single-word frequency list alone is not sufficient and must be supplemented by psychological considerations. In addition it must be weighed alongside the frequency of fixed expressions, many of which will equal or exceed in frequency the single-word forms.

Vocabulary in use

One characteristic of words at this advanced level has been mentioned: their proclivity to generate metaphorical meanings. Another characteristic is their tendency to have connotations and degrees of nuance and subtlety which the core 2,000 words are able to operate independently of: words like *table, hand, blue, cup, water*, etc. can be learned with their basic core meanings at the elementary level and it would be considered wasteful of time to dwell on their possible cultural or obscure connotations (e.g. *blue mood* or *blue pencil* [the latter referring to censorship]). Words in the 6,000 to 10,000 word band seem less capable of innocent use, and much focus will need to be on the connotations of words in their typical contexts of occurrence, as well as grappling with their meaning. Take the expression *peace and quiet* (18 occurrences in 5 million words), already mentioned above in the context of fixed expressions that will present themselves as candidates for inclusion in the advanced level vocabulary syllabus owing to their salience and their level of frequency which, though it may place them outside the 30+ level of the 6,000-10,000 word band, is high enough to be noticeable and to generate observable (and useful and teachable) patterns of usage. Figure 4 shows a key-word-in-context (KWIC) concordance for *peace and quiet*. We note it is not neutral, but typically associated with contrastive contexts, where someone needs or finds peace and tranquillity in contrast to some other (negative) situation where noise or lack of peace and tranquillity is/has been problematic:

1 recognise the need for a little visual *peace and quiet* occasionally.
2 ning every skirmish on the streets. For *peace and quiet* the walker
3 ded to share a vacation in the relative *peace and quiet* of Beirut.
4 only contacted the police to get some *peace and quiet* because her
5 men who wish to while away the hours in *peace and quiet* with a rod
6 of-term exams to study for an' I need *peace and quiet* for a while.
7 by the possibilities they offered for *peace and quiet* writing to
8 Price used to come here for a bit of *peace and quiet* Tom remarked
9 It is the penalty, perhaps, for such *peace and quiet*. Some years
10 as I can have my beer and eggs in *peace and quiet*. He looked
11, charming beaches and countryside, and *peace and quiet*. And the dog

12 all we wanted was a bit of *peace and quiet*. He didn't
13 resort 18 months ago hoping to find *peace and quiet*. Instead she.
14 when she was having a bit of *peace and quiet*. She had on
15 London for a while to convalesce in *peace and quiet*. Sean felt a
161 yourself with a long-term poultice of *peace and quiet*. Squadron-
17 we did nothing, Inspector. We wanted *peace and quiet*. We had no
18 treatments, exercise classes and *peace and quiet*. She found

Figure 4: Concordance for peace and quiet (5m words written)

Thus in the case of wanting to make, for example, a neutral statement that one loves to live in the country because it is peaceful and tranquil, *peace and quiet* may not be appropriate, implying as it does a contrast which the speaker/writer may have no intention of making. This is typical of the lexical issues that have to be tackled at this advanced level, since the connotations of words seem to emerge more powerfully. Other expressions within the idiom band of 10-30 occurrences which have significant connotations include *(jump) on the bandwagon, bring/come to a head* and *have second thoughts*. Alexander (1985) sees the problem of phraseological knowledge as one of the key issues in learning and using vocabulary at the advanced level, and notes that for metaphorical idioms, the kind of knowledge needed is overlaid by cultural connotations.

Use also includes collocations, the unpredictable, probabilistic combinations of words that most often simply have to be learnt in any second language. Collocation bears a relationship with frequency such that lower frequency words tend to collocate more strongly, thus inappropriate collocations are likely to be more easily made with lower frequency words from the learner's viewpoint, owing to lack of exposure. The verb *generate* (135 occurrences) is a typical case in point. Apart from its intuitive collocability with *electricity/energy*, it displays a limited set of environments which includes financial/money-related concepts, feelings and emotions, and, more recently, computer applications. Figure 5, a random sample concordance of 30 lines illustrates this (collocations in bold):

1 Labour's early months in office *generated* an enormous **sense of excitement**
2 The Cycleway is expected to *generate* **an extra £18m** each year through
3 and dub. It's a big place to *generate* **atmosphere** so make sure it goes
4 atively small investments could *generate* big savings in current spending
5 NetBenefit. To see **statistics** *generated* by Analog 2.0 for this **server**
6 Much of this **traffic** is in fact *generated* by countryside attractions suc
7 of the **traffic** on our roads is *generated* by people travelling to and fr
8 urprise. The extensive **computer-** *generated* effects, added after shooting,

```
9  and steam coal, mainly used to generate  electricity. Coking coal is a y
10 countries use nuclear fuels to generate  electricity. Some people are wo
11 hey put acid in metal cases to generate  electricity. Wow man, basic!
12 from their movement is used to generate  electricity; (generator, the du
13  of a bike dynamo? How does it generate  electricity? Explain how a p
14  ilst the economy struggles to generate  enough income for the populatio
15  nth since re-opening. We have generated enough money to get us through
16  re over the bank's ability to generate  enough profits quickly enough t
17   local councils, Richmond, to generate  extra cash by acting as an inte
18  ayered digital plan of London generated from aerial photography - is d
19    Republic. The deal does not generate  goodwill, which would have to b
20   His reappearance would still generate  great media interest, although
21  leading exponents of computer-generated imagery. The model - a multi-l
22  ilm, photography and computer generated images. Regional Time Pl
23  o play attractive football to generate  profits. The same old footba
24  t the Americans have suddenly generated publicity about the share valu
25  up now.'The advertisement has generated quite a lot of interest and an
26  dging that extra growth would generate  revenues on top of this by cutt
27   place there if only we could generate  the cash and energy to make it
28  th the income that my capital generated. The trust, administered by Pe
39 tional way.  The only computer-generated version of the old superhero i
30  ship project designed to help generate  widespread public enthusiasm f
```

Figure 5: Concordance sample for generate (CIC 5m words written)

Collocations will therefore be a major aspect of advanced level teaching, and learners may have to be explicitly introduced to the importance of the notion of collocation via awareness-raising activites, since many learners, even at advanced level, see vocabulary-learning as largely a matter of confronting single words. We may indeed conclude that collocations, along with idioms, form the main component of the multi-word lexicon and that the multi-word lexicon is at the heart of advanced level lexical knowledge.

Conclusion

We are now in a position to revisit the aims of an advanced vocabulary learning syllabus sketched out in the beginning of the paper. Our corpus-based investigation has partly given us answers (to the quantitative questions) and partly suggested directions and guidelines to the more qualitative issues. Let us remind ourselves of the aims:

- To push the vocabulary size towards comprehension targets above 90% for typical texts. This seems feasible, and involves aiming for a 10,000-word receptive knowledge. The advanced learner can be expected to come to the task with anything from 4-6,000 words already known, presenting a learning target of around 4-5,000 words to achieve good, fluent reading levels. Most teachers will recognise, however, that 5,000 words is an impossible target for classroom teaching as such, and its achievement will depend on motivated work out of class, including extensive L2 reading and awareness of learning strategies which will be available both during and after formal/institutional learning.

- To expose the learners to a vocabulary that corresponds to frequency levels beyond the first 6,000-word band, but which is not too obscure to be tantamount to useless. Corpus-based techniques come into their own in this domain, since, even at lower levels of frequency, it is possible to generate word-lists which differentiate low frequency items from rare items. The major proviso on this issue is related to the mismatch between observed frequency of use and the powerful, natural tendency of the mind to learn associated sets of items which can be retrieved as wholes, as well as the notion of psychological saliency, which may provide the curiosity and motivation to learn even rare items such as idioms. And on this last point, it was noted that the corpus size may need to be expanded in order to generate sufficient occurrences of salient but infrequent items so that relevant patterns of use can be seen to emerge.

- To impart the kinds of knowledge essential for using words at this level of sophistication, given their specific lexical content. We noted that words at the advanced level, because of their low frequency, tended to bring the challenge of more extended meanings and more obvious cultural connotations with them, in the sense that the high frequency words can be and usually are dealt with at lower learner levels in terms of only their core, basic meanings, an eminently sensible way of tackling the polysemic nature of most words in graded learning, in the view of Lennon (1990). Connotations and recurrent collocations can be traced using concordances. Here also the issue of extended chunks come into play, with idioms being part of the desired learning load, and corresponding questions about their use and distribution (e.g. in speech versus in writing) being to the fore.

- To develop awareness and skills that will stand the learner in good stead for becoming an autonomous vocabulary-learner. This is a question of developing activities alongside the actual learning of words which introduce to the learner notions such as collocation, metaphor, etc. For example, many learners have an awareness of idioms of the *verb+complement* type (*hit* *the sack, carry the can, jump on the bandwagon*), but how many learners are aware of the pervasiveness in everyday language of binomial idioms (*rough and ready, part and parcel, out and about, down and out*)? Explicit focus on such items may be necessary to tune the learner's antennae to be receptive to new ones when they are used both in and out of class. Lexical skills include ways of maximising learning opportunities during interaction (e.g. asking for paraphrases, probing the meaning of unfamiliar items with one's interlocutor, etc.). Above all will be developing the awareness that the class or textbook will only scratch the surface of the vast store of low frequency vocabulary and that the onus will be on the learner him/herself to achieve the final target, but that the target is achievable given the right strategies and motivations.

In sum, the advanced level vocabulary programme need not be a haphazard free-for-all where planning and organisation simply dissolve into the fog of 50,000 unknown words. With a combination of corpus-based research and strategic training for learners who will have to, if we are honest, complete the task for themselves, we can go at least some way towards presenting an advanced level programme worthy of the word *programme*. And that's a loaded word.

Notes

[1] The figure of 2000 words is often considered a benchmark not just for speaking, but as a vital threshold in reading, including reading academic texts with the addition of a core of academic words (Nation, 2001:17).

References

Alexander, R. (1985). Phraseological and pragmatic deficits in advanced learners of English: problems of vocabulary learning? *Die Neueren Sprachen* 84(6): 613-621.

Carroll, J. B., P. Davies and B. Richman (1971). *The American Heritage Word Frequency Book*. New York: Houghton Mifflin.

Carver, R. (1994). Percentage of unknown vocabulary words in text as a function of the relative difficulty of the text: implications for instruction. *Journal of Reading Behavior* 26: 413-437.

Honeyfield, J. (1977). Word frequency and the importance of context in vocabulary learning. *RELC Journal* 8(2): 35-42.

Hu, M. and I. S. P. Nation (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language* 13(1). (On-line journal at: http://nflrc.hawaii.edu/rfl/)

Lennon, P. (1990). The bases for vocabulary teaching at the advanced level. *ITL Review of Applied Linguistics* 87-88: 1-22.

McCarthy, M. J. (1998). *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

McCarthy, M. J. (1999). What constitutes a basic vocabulary for spoken communication? *SELL* (*Studies in English Language and Linguistics*), 1: 233-249.

McCarthy, M. J. and F. O'Dell (1999). *English Vocabulary in Use (Elementary)*. Cambridge: Cambridge University Press.

McCarthy, M. J. and F. O'Dell (2001). *Basic Vocabulary in Use*. New York: Cambridge University Press.

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

Richards, J. C. (1974). Word lists: problems and prospects. *RELC Journal* 5(2): 69-84.

Using corpora to teach and assess vocabulary

Norbert Schmitt
University ofNottingham

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.

Introduction

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 (e.g. *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
              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
                         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,
      I mean they can wear if you think about it they wear suits in the
      And why, they don't need to think about it, they can talk you out of

   enetrating as lasers. 'We might think about that, ' I say at last.
          I'll have to start and think about that train, Dwight.
        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
```

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).

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
 rying 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, 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:

$$\text{Type/token Ratio} = \frac{\text{number of separate words (types)}}{\text{total number of words in the text (tokens)}} \text{x}100$$

(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 this is affecting both the content and tone of learners' 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, influx, 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 more 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, 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.

Notes:

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

a. The RANGE and WORD vocabulary analysis programs by Paul Nation are available free of charge at : http://www.vuw.ac.nz/lals/

b. 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/

c. Another concordancing package is MonoConc Pro, which is available at: http://www.athel.com

2. 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 Guide to the Computer Analysis of Language*. Edinburgh: Edinburgh University Press.

Carter, R. and M. McCarthy (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 (1): 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 T. Odlin (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 I. S. P. Nation (1995). Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics* 16 (4): 307-322.

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

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

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

Schonell, F., I. Meddleton, B. Shaw, M. Routh, D. Popham, G. Gill, G. Mackrell and C. Stephens (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: key 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., D. Schmitt and C. Clapham (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing* 18: 55-88.

Simpson, R. and J. Swales (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 G. E. Perren and J. L. M. Trim (eds), *Applications of Linguistics: Selected papers of the Second International Congress of Applied Linguistics, Cambridge, 1969*. Cambridge: Cambridge University Press.

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

Fixed expressions, prepositional clusters and language teaching

Melinda Tan
Assumption University, Bangkok.

Abstract

This paper attempts to bring a new contribution to knowledge about fixed idiomatic expressions in English, by demonstrating that the overall meaning of such expressions need not always be conveyed by the presence of lexical words. Linguistic observation of natural and authentic language use has shown the existence of fixed idiomatic expressions consisting solely of grammatical words and possessing a particular overall meaning. The paper will illustrate through the use of corpus evidence the existence of a particular type of fixed expressions called prepositional clusters which are commonly found in everyday informal communication (written and spoken). Examples of prepositional clusters are "round and round", "ups and downs", "on and off", etc. Applications of corpus linguistic principles in the classroom will also be discussed through the illustration of some possible activities which can be used to teach idiomatic expressions like prepositional clusters. These activities aim to develop in learners skills of Noticing, Hypothesising and Experimenting in order for them to become sensitive to how patterns of language convey meaning usages and grammatical functions.

Fixed expressions and grammatical words

Corpus studies and other works related to fixed expressions, idiomaticity and metaphoricity have attempted quite successfully to isolate, describe and classify huge numbers of conventionalised utterances formally, semantically and pragmatically. A vast majority of these studies however have focused on fixed expressions composed of lexical words, thus endorsing the widely held view that conceptual, idiomatic or metaphorical meaning can only be conveyed through lexical words. However, everyday observation about natural language use shows this notion to be only partially true. Idiomatic and metaphorical meaning can also be expressed by fixed expressions consisting solely of grammatical words. I will make use of Stubbs' (1986a) definition of the difference between grammatical and lexical words. Thus:

> "Lexical words are nouns, main verbs, adjectives and adverbs. Grammatical words are anything else: pronouns, conjunctions, articles, prepositions, auxiliary and modal verbs. There are many tests to distinguish these two classes, but very briefly it can be stated that lexical words comprise large open sets with hundreds and thousands of members in common use, whereas grammatical words comprise small closed classes with only a few (less than around 20) items in common use"(Stubbs, 1986: 27)

Examples of fixed expressions composed solely of grammatical words from the following categories are:

a) prepositions or phrases derived from words which function also as prepositions e.g. *in and out, ins and outs, up and down, ups and downs, over and beyond, round and round, inside out, in for, out to, etc*

b) demonstrative pronouns e.g. *this and/or that*

c) adverbs or phrases composed of words which function also as adverbs e.g. *here and/or there, now and then, now and again, above and below, etc*

d) conjunctions e.g. *either..or, neither...nor*

It is also possible to have fixed expressions composed from combinations of grammatical categories such as:

a) adverb or adjective + preposition e.g. *all for, much of, except for, etc*

b) adverb + adverb e.g. *very much, very little, much more, etc*

c) adverb + conjunction e.g. *all but, in that, etc*

d) preposition + adjective e.g. *in all, etc*

(All the words listed in the examples above are considered grammatical according to Stubbs' (1986:27) Function Word List).

The fixed expressions given above are commonly found in both written and spoken English. Their formation as a result of combination with other grammatical words creates a fixedness in structure and even an idiomatic or metaphorical meaning. However, there are many other examples of fixed expressions found in authentic language combination. Below are examples taken from a concordance search which reveals the innumerable fixed expressions which are found in English, composed of prepositions, which are grammatical words according to Quirk *et al.'s* (1985), Stubbs' (1986a, 1986b), Finocchiaro and Brumfit's (1983) as well as Carter's (1998b) classification of grammatical categories. The examples below are taken from the tagged CANCODE[1] corpus.

```
sHave] got [VPpast] one [M]. In [T] and [Cand] around [T] the [Dt
n= local [Jbas] venues [Npl] in [T] and [Cand] on [T] the [Dthe]
FpastHave] trekking [VPpres] to [T] and [Cand] from [T] the [Dthe
und [A] thirty two to thirty on [T] and [Cand] off [T] thirty [M]
orward [A]+ Right [VI]. +during [T] and [Cand] after [T] the [Dth
] of [T] communication [Nsg] to [T] and [Cand] from [T] Eastern [
```

[A]=Adverb, **[Cand]**=and, conjunction, **[Nsg]**=Noun, singular, **[Npl]**=Noun, plural, **[Jbas]**=Adjective, base, **[Jcomp]**=Adjective, comparative, **[Jsup]**=Adjective, superlative, **[Nsg]**=Noun, singular, **[Npl]**=Noun, plural, **[T]**=preposition

```
[VFpast] back [A] there [A] for [T] about [A] three [M] or [Cand]
f [T] door opens split [Nsg] in [T] about [A] fifteen [M] differe
VFpast] in [T] price [Nsg] from [T] about [A] two [M] pound [Nsg]
o [M] pound [Nsg] twenty [M] to [T] about [A] three [M] quid [Npl6679

Da] bit [Nsg] cold [Nsg] . +for [T] about [T] another [Dind] thir
fly [VI] to [T] Dublin [Nsg] in [T] about [T] half [Dpre] an [Da]
stHave] the [Dthe] lot [Nsg] up [T] in [T] the [Dthe] north [Nsg]

[VFpresBe] backing [VPpres] out [A] of [T] this [Pdem]. It's er Y
j]. We [Ppers] went [VFpast] in [A] to [T] the [Dthe] little [Jba

[Ppers] 'll [VFmod] end [VI] up [A] with [T] broken [VPpast] legs
VFpast] a [Da] runner [Nsg] out [A] of [T] the [Dthe] door [Nsg].
n [VFpres] sort [Nsg] of [T] in [A] between [T] jobs [Npl] and [C
ries [Npl] going [VPpres] round [A] about [T] erm [Aintj] people
```

Prepositional Clusters as an object of study

The analysis that follows will focus on one type of idiomatic fixed expression that is composed of grammatical words: prepositional clusters. A prepositional cluster can be defined as one which is a binomial or a compound that contains only prepositional constituents.

It must be noted that English prepositions are a notoriously difficult area to teach due to their anomalous nature with regard to meaning. It is probably for this reason that the teaching and learning of prepositions have traditionally focused on the various meanings related to a single preposition rather than the binomial or compound unit. According to Rastall (1994), focusing on the single prepositional unit is the simplest and least problematic approach to the teaching of prepositions:

> "...so long as we restrict ourselves to the expression of simple spatial relations and movements, the teaching of prepositions in English presents relatively few problems..." (Rastall 1994: 229)

Selecting prepositional clusters for analysis

A frequency table of a tagged corpus (CANCODE), shown in Frequency Table 1, revealed that the most common grammatical class of collocates that co-occurred with a preposition were another preposition [T] and AND which occured in positions 14 and 15 respectively. The only other grammatical collocate present was THE which did not interest me much because it is a determiner which would immediately follow after a preposition according to English grammar rules and is thus unlikely to form any fixed expression.

Frequency Table 1: Collocates for Prep [T])

WordSmith Tools Collocates

frequency... based on 16000 concordance entries

1. NSG	14188
2. A	6782
3. PPERS	5435
4. THE	5007
5. OF	4705
6. THE	4562
7. IN	3554
8. TO	3208
9. NPL	3106
10. AINTJ	3008
11. JBAS	2476
12. VI	2377
13. DA	2370
14. T	2134
15. AND	2043
16. IT	1778
17. VPPRES	1688
18. CAND	1685
19. YOU	1574
20. M	1512

Notes:

A = Adverb **AINTJ** = Adverb, interjection **CAND** = Conjunction

DA = Indefinite article **DTHE** = Definite article **JBAS** = Adjective, base

M = Number **NSG** = Noun, singular **NPL**= Noun, plural

PPERS = Pronoun, Personal **T** = Preposition **VI** = Verb, Infinitive

VPPAST = Verb, participle, past **VPPRES** = Verb, participle, present

Using the collocates [T] (preposition) and AND as examples of grammatical constituents in a prepositional cluster, a possible syntactic pattern was found: **[T] + [T] (Preposition + Preposition).** Many examples of the above cluster are commonly found in written or spoken English such as *round about, down under, inside out, etc.*

The second pattern, [T] + AND was initially attempted. However, since there are infinite collocational possibilities with this combination, the number of collocate possibilities was narrowed when the pattern was restricted to the node {[T] + AND}. Consequently, on observing the collocates that co-occurred within a span of four words to the left and right of the node {[T] + AND}, it was found that a common grammatical collocate that immediately followed after the node was yet another preposition [T] (see Frequency Table 2 below). The second cluster pattern which was then formed was **{[T] + AND + [T]} (Preposition + AND + Preposition)** since there are many common prepositional clusters in English that exhibit this particular pattern. Some examples of these clusters are ***in and out, ins and outs, by and by, on and on, over and over, etc*** which are used frequently in written and spoken English.

Frequency Table 2: Collocates for Prep + And ([T] + AND)

WordSmith Tools Collocates

frequency... based on 272 concordance entries

1. AND	296
2. CAND	285
3. TO	129
4. PPERS	110
5. A	101
6. NSG	70
7. VI	70
8. WITH	50
9. IN	42
10. VPPAST	42
11. VFPRES	42
12. I	37
13. THE	36
14. FOR	34
15. VFPAST	34
16. VPPRES	34
17. IT	33
18. YOU	33
19. T	30
20. OF	27

Prepositional Clusters as lexical units of meaning

Data taken from three corpora, the BNC sampler (written and spoken, 50 examples), COBUILD (written and spoken, 40 lines) and CANCODE (spoken, 40 lines) indicated that prepositional clusters are lexical units of meaning in two respects:

a) meaning usage: a fixed prepositional cluster could have a meaning usage, different from its components. Some of these meaning usages are metaphorical.

b) grammatical distribution: a fixed prepositional cluster could have less varied grammatical functions than its components

Analysis 1: Prep + Prep: round about

Comparison of grammatical behaviour between the cluster **round about** and its components **round** and **about**

On close analysis, it was found that as a cluster, *round about* occurs **mainly** in adverbial position indicating place (orientation, direction) and time (when) whereas its components *round* and *about* could occur as either prepositions or adverbs, thus already marking the cluster's formal difference grammatically from its components. The data below will demonstrate this difference.

Components: *round, about* (prepositions)

...black curls bunching *round* the rim of his hard hair...

...clawed his way *round* the car...

...received the heart of a lad *round* the corner...

...there are a lot of myths *about* babies

...I fee pretty good *about* our government...

.....Telling you *about* his experience, you know...

Components: *round, about* (adverbs)

...and screaming in anguish as the nose slithers *round*, that the Dodge...

...hitting the wall, sickening thuds, cracked his head, turning *round* and *round*...

...Stephen brought Bill *round* and we spent an amusing...

..some models are still *about,* alive and kicking...

...Cromwell knocked it *about* during the civil war...

...well...*about* four hours ago...

Cluster: *round about* (adverb modifier of place-orientation, direction)

...And that's in that place just *round about there....*

...we returned by the water side *round about the North-point...*

Cluster: *round about* (adverb modifier of time - when)

...the post-war pattern settled into what looked like immobility *round about 1950...*

Round about four o'clock in the afternoon, he would sometimes forget Morris....

...*Round about the same time*, Douglas...

...It's going *round about 250 earth days...*

...when was it dear ? - *round about two ...*

...you're looking at about *round about twenty years...*

...actually have a meeting *round about early autumn...*

...I think it was *round about the time* I started seeing...

While the above examples demonstrate that *round about* functions mainly as an adverb showing an estimation of place and time, most grammar coursebooks books have traditionally tended to regard its components, *round* and *about* as single prepositions only. Also their usage is usually explained in terms of place and subject matter respectively as shown in the following which have been taken from popular EFL coursebooks:

- *Mr Wood has brought the car* ***round*** *the house* (taken from English Practice Grammar, 1995, pg 178)

- *The bus-stop is* ***round*** *the corner* (taken from Essential Grammar in Use, (Elementary) 2nd Edition, 1997, pg 214)

- *We talked* ***about*** *a lot of things in the meeting* (taken from Essential Grammar in Use (Intermediate), 2nd Edition, 1994, pg 264)

The above three sentences illustrate that this is an erroneous treatment of prepositions in general, because a preposition like a*bout* is used not only to mark a relationship between two entities grammatically but more frequently, when it is in combination with *round* conveys the meaning of approximation. I shall consider the two sentences:

a) *I'll come* ***about*** *four o' clock*

b) *I'll come* ***round about*** *four o' clock*

Both sentences can be observed to mean the same as the sentence "I'll come *approximately* at four o'clock". However, most of the time, examples of the type found in sentences a) and b) are not taught in EFL classrooms to convey the meaning of approximation but more traditionally *about* is taught as a preposition related to subject matter e.g. "Tell me *about* your adventures". In short, the predominance of examples illustrating only the deictic usage of *round* and *about* and the absence of fixed expressions composed of prepositional clusters e.g. *round about* (including many others) illustrates that coursebook writers have tended to rely on intuition rather than observation of authentic language usage.

The following example "*He told me about his new adventures*" taken from a grammar coursebook (A University Grammar of English, 1987: 45*)* illustrates the kind of prescriptive pedagogy more commonly involved in dealing with a preposition. In the coursebook, *about* is taught as a complementation of a verb or adjective" (Quirk and Greenbaum, 1987: 45 in A University Grammar of English).

Furthermore, when trying to convey the sense of time, it is commonly found that EFL students are taught to use the preposition "at" to speak about time - "I'll come *at* four o'clock" - rather than "I'll come *round about* four o'clock". However, data from the corpora shows that *about* and *round about* are commonly used to convey the meaning of approximation of **both** time and place references in informal spoken English rather than "at". The conveyance of the sense of approximation is also supported by data from corpora such as those below:

...it will stand me in good stead when I'm *about* 40 and ....

...and the actual figure would be *about* 13%...

...it took them *about* five minutes of hemming and hawing...

...*round about* the time the baby's born...

In the case of the preposition *round*, it is interesting to note that while grammatically marking a relation between two entities, it behaves syntactically as synonymous to the other preposition *around.* Both can be used interchangeably as prepositions in the examples "I live round here" and "I live around here". However, in comparison to the cluster *round about*, there were no instances of data showing the composite *round* conveying the meaning of approximation. Rather it emerged that it is used as a preposition, adverb, phrasal verb, adjective and noun to convey a sense of orientation or a meaning of circularity. The following sentences demonstrate this observation:

preposition:

....we went right *round* the grounds...

...followed Lorenzo *round* the sides of the church...

...began to make a slip-knot *round* it...

adverb:

...and then I turned *round* and saw the telly...

...gently work it *round* in larger circles...

...coax his niece *round* to his point of view...

phrasal verb:

...and *round off* with a piece of Lemon Madeira...

adjective:

...and pale *round* black rocks...

...addressed in my mother's *round* handwriting...

noun:

...three holes of his third *round*?

...Second *round* of elections...

The above examples demonstrate that the composite *round* can take on numerous grammatical functions unlike the cluster *round about,* which serves a mainly adverbial function. Contrary to intuition, clusters do not serve exactly the same grammatical function as their components. This observation has implications for language teaching because it establishes the need to consider how students can be made critically aware of such differences in grammatical behaviour between the two. EFL learners of English should not be misled into assuming through logical reasoning that knowledge of the grammatical functions of certain word forms would automatically apply to a cluster that is composed of these word forms, as evidence from the corpora shows otherwise. It is to prevent such misconceptions that I agree with one of Michael Lewis' (1997) suggestions that a way of teaching language could be following a slot-and-filler framework, similar to the word pattern framework proposed by Hunston, Francis and Manning (1997). This framework would be better equipped to allow a more holistic development of a learner's lexico-grammatical knowledge through a critical awareness of how grammar and vocabulary are inextricably bound in many ways and are not two linear and parallel aspects of language learning .

Differences in meaning usage between *round about* and its *components*

While the previous section demonstrated quite clearly that a prepositional cluster could have a linguistic identity different from its components by virtue of its less varied grammatical functions, this claim about the cluster having its own linguistic identity has not been supported in coursebooks. A popular grammar coursebook like *Collins Cobuild Grammar of English* (1995) has also stated that "many words can be used as prepositions and as adverbs with no difference in meaning...". This statement is misleading because it seems to exclude the existence of phrasal or multi-word units which are composed solely of grammatical words and are commonly found in written and spoken English. Thus the statement presumes that a random selection of two prepositions combined together in a fixed cluster like *round about* would have "no difference in meaning" as the individual constituents (*round, about)* of the cluster. However, the next few sections will seek to demonstrate, using the example *round about,* that the components *round* and *about* convey somewhat different meanings from the cluster *round about.*

Cluster:* *round about

Sense 1: Approximation

...to manage a husband and six children in three rooms on *round about* a pound a week...

...where she starts crying *round about* the end of the queue for tickets...

...No, I said *round about* six o one

...it's *round about*, it's *round about* the same...

...the early 1950's, *round about* then...

...the suburbs of places *round about* Jerusalem...

...comes *round about* half four...

...its usually *round about* mid January...

In comparison, *round* showed commonly meanings associated with circularity and orientation whilst *about* showed a meaning related to subject matter. It is thus inaccurate to assume that *round about* as a fixed cluster has *only* one meaning related to its components and no other, as this has been shown not to be the case from previous examples of *round* and *about* taken individually. In fact from analysis of the data, it was found that *round about* had **five** other different meanings besides "approximation". These other meanings were "indirect", "road junction" (when *round-about* is hyphenated)"surround to protect", "in the vicinity" and "concerning".

Sense 2: "circuitous"

The meaning of *round about* in this case to mean "indirect" was found in the following data:

...the melody flowed up and down and *round about* in a long cadence...

...She led Gwer up by a *round about* way, then waited...

...You could have spoken in a very *round about* way...

...Kind of a *round about* cousin...

...They said it in a *round about* way for two and a half hours...

...this sort of longer *round about* ways...

Sense 3: Traffic Road junction round central island

By figurative and literal extension of the previous meaning of "indirect", it is clear how the meaning *round-about* (hyphenated so as to function as a noun) as "traffic road junction round central island" was derived.

...just after the *round-about* intersection with the B3274...

The reference to the words "intersection" and highway number "B3274" indicates its common use in motoring.

Sense 4: Surround to protect

The meaning of "surround to protect" is demonstrated below:

...and *round about* him, his band of assorted...

...a good mud wall to be cast up *round about* the factory...

...built of cedar and fortified *round about* with sharp trees...

Sense 5: in the vicinity

The disambiguation of this particular meaning was observed after analysing the following data:

...it's based on villages and towns *round about*, like Barancija...

...we visit the little villages *round about*...

...for the town and all the farmers *round about*...

...the people *round about* hissed and told her to sit down...

...I glanced *round about* myself, on the lookout for clues...

...places to stay *round about* where they were...

The discoursal function of *round about*

Analysis of the data also revealed an interesting discoursal function of the cluster *round about*, which is the conveyance of vagueness or ambiguity on the part of a speaker or writer (see Channell, 1994). Analysis showed that facts, be it the actual time, place, location or direction where events or actions had taken place, tended to remain obscure and indeterminate on the part of the speaker or writer. This observation was discerned from the following examples:

...Who remembers the public-service training films they used to show on TV, *round about news time*

...the ones that start just above the knees and peter out somewhere *round about the coccyx*

...as I was saying usually between sort of *round about the middle of the day* there's people knocking on the door

...I lost interest *round about week two*

...between the two stations *round about 500 times a second*

...the charismatic renewal system *round about 1973* was the most notable

...usually *round about mid January*

Summary of differences between *round about* and its *components*

The table shows diagrammatically the difference between the cluster *round about* and its components *round* and *about*. It can be seen that the cluster does not correspond similarly to its components in the aspects of grammar and in meaning usage since the grammatical function as well as the distribution of meaning associated with the components and cluster are different. This observation corresponds to Sinclair's unit of meaning as "a single, independent meaningful choice of words normally showing independent variation" and "can be associated with a distinct formal patterning" and (See Sinclair, 1991a:6 and 1996: 75) which implies that the cluster *round about* would qualify as a single lexical unit with its own lexicogrammatical environment.

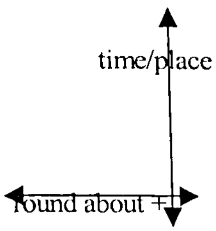
Grammatical functions	**round about (cluster)**	**round (composite 1)**	**about (composite 2)**
Preposition	♣	♣	♣
Adverb	♣	♣	♣
Noun			
Adjective			
meaning usages			
circularity		♣	
course of action		♣	
subject matter			♣
approximation	♣		♣
surround	♣		
vicinity	♣	♣	
indirect path	♣		♣
traffic junction	♣		

Paradigmatic and syntagmatic differences between *round about* and its *components*

Since the cluster *round about* differs from its components *round* and *about* in terms of grammatical function and usage, this would imply thus that paradigmatically and syntagmatically, the cluster and its components would prospect for their own entries from a pool of potential lexical words; content or functional (see Tognini-Bonelli, 1996).

A) The cluster: **round about**

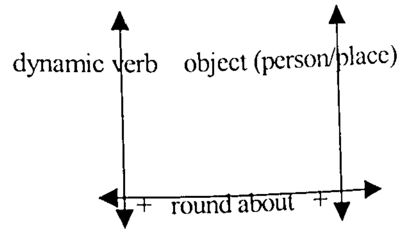
1) **Approximation**



e.gs. places *round about Jerusalem*

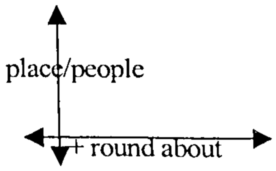
come *round about half four*

2) **Surround**



e.g. cast up *round about the factory*

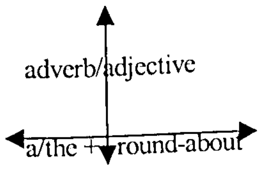
3) **vicinity**



e.gs. the little village *round about* here

people *round about* her hissed

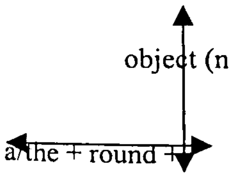
4) **traffic-junction**



e.g. the huge *round-about* intersection

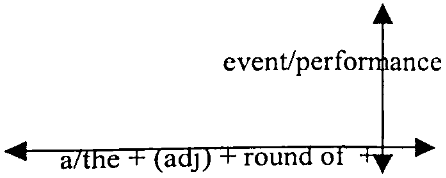
B) The composite *round:*

1) **circularity**



e.gs. the '*round*' dance

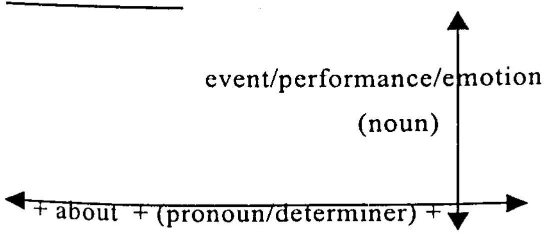
2) **course of action**



e.g. a tough *round* of negotiations

C) The composite: *about*

1) **subject matter**



e.g. to talk *about* their grief

The diagrams above illustrate that the lexical choices on the paradigmatic axis of the cluster *round about* and its components *round* and *about* similar for each derived meaning of the cluster and those of its individual components. Furthermore, the types of word classes that the lexical choices prospect by virtue of their lexis is different in each case. Thus by observing the interrelation of lexis and grammar, it can be shown once again that the cluster and each of its components should be considered individual units of meaning.

The findings for prepositional clusters *in and out, ins and outs* showed similar indications that prepositional clusters were lexical units of meaning in their own rights. The table below indicates these findings.

grammatical function	**cluster 1 "in and out"**	**cluster 2 "ins and outs"**	**component "in"**	**component "out"**
Adverb	♣		♣	♣
Adjective			♣	
Noun		♣		
Preposition			♣	♣
Phrasal verb			♣	♣
As part of many fixed expressions			♣	♣
meaning usages				
repeated action	♣			
intricacies, complexities and details		♣		
inclusion within				
time			♣	
space			♣	
within a circumstance			♣	
within a particular sphere or field			♣	
expressing a particular state or emotion			♣	
fashionable			♣	
movement to the exterior				♣
exclusion/dismissal				♣
not in fashion				♣
discover and examine				♣
not in use				♣
attribute or part of a collection/ organisation				♣
distribution				♣
extinguished				♣
not functioning				♣
defeated				♣
not in power				♣
motivation				♣

Teaching prepositional clusters in the classroom

This section will demonstrate how prepositional clusters can be taught in the classroom through the use of syntactic patterning. The focus here is to develop an awareness of language use and patterning through three skills:

noticing, hypothesising and experimenting. These three skills constitute what I call an Investigative-Oriented Approach (IOL). Below are two sample tasks with commentaries explaining how the tasks can be taught to learners of different linguistic levels.

TASK 1:

A) Prep+and+Prep: *e.g. ups and downs, up and down*

Read the following extracts and **guess the meanings** of *ups and downs* and *up and down.* What is the **grammatical function** of both ?

- "To be fair, he tried to understand, but he (like most men) wanted a relationship similar to his parents', where the woman would be there for him in the evenings and prepare his dinner, listen to the *ups and downs* of his day. When I wasn't back, or I was bent over a computer piece, he felt unloved and unwanted..."
- "We decided to forget about the third canister and made off across the cornfield at speed, the jeep bouncing *up and down* on the very uneven surface..."
- "The trains, running *up and down* from London to Stanmore and back, could only be seen through the foliage as a series of silver flashes, but their singing rattle made a constant background music..."
- "We hate to detain our most welcome guests, especially when they have come from so far. He looked Anna May *up and down* as if his mind could do with a good Chinese laundering. I have urgent business to attend to..."

This is a simple task which is suitable for beginner to intermediate learners. Students are encouraged to make use of the skills of noticing and hypothesising. They start by observing how the common idiomatic expressions *up and down* and *ups and downs* differ from one another in terms of meaning usages and grammatical functions. In order to do this, they have to observe the kinds of verbs that collocate with each of the idiomatic expressions as well as make use of their knowledge about grammar to observe the grammatical functions of each. Students then make use of their observations to hypothesise about the meaning usages and grammatical functions.

TASK 2:

B) Prep + Prep: e.g. *round about*

Some of the meanings of the prepositional cluster *round about* are given below, together with their grammatical constructions which make up the meanings.

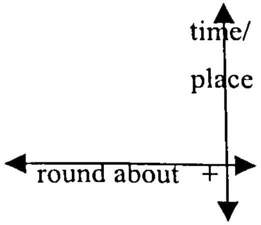
Meaning 1: round about (to show approximation in time and location)

e.gs. *Round about four o'clock in the afternoon,* he would sometimes forget Morris...

...we returned by the water side *round about the North- point...*

Representation of Form and Meaning:

1) Approximation

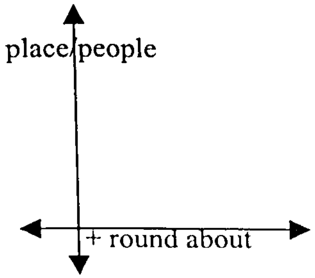


Meaning 2: round about (vicinity)

e.g. ...there are places to *stay round about where they were.*

Representation of Form and Meaning

2) vicinity

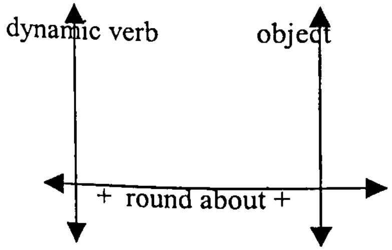


Meaning 3: round about (surround and protect)

e.g. ...a good mud wall to be cast up *round about the factory*

Representation of Form and Meaning

3) Surround



Question:

There is 1 other meaning of *round about.* Try and guess its meaning from the data given and draw a diagrammatic representation of form and meaning, to demonstrate its usage.

...turn left after the *round-about* at the intersection...

...just after *the round-about intersection* with the B3274

This task is more suited for upper-intermediate to advanced learners. The skills involved here are noticing, hypothesising and experimenting. Students first have to observe how the various meanings of *round about* have different collocating verbs and objects. Then they have to construct hypotheses, from the given examples, about how each particular meaning of *round about* has a special colligating pattern and specific semantic preferences for particular words to create positive or negative prosodies in their usage. To practise the skills of noticing and hypothesising, students can experiment with these skills in the final question of the task, in which they are required to draw a new form-meaning representation for a particular meaning of *round about.*

Applicability of IOL tasks in the classroom

This paper will conclude by describing how IOL tasks show characteristics that make them suitable for application in the classroom: **information gap, feedback, evaluation, authentic materials and transfer of learning.**

Information gap can be defined as the information that one person in the exchange knows but not the other. In IOL, an information gap exists because the students in the class do not know the answers to the task beforehand and can only find out the answers through a process of investigation.

IOL confines itself simply to the development of three skills - Noticing and Hypothesising and Experimenting - and is not concerned with the development of fluency or accuracy in communication. The tasks are monothematic in nature and consist solely of analysing various examples of common language patterns, in order to investigate their usage. **Evaluation** is thus measured according to the extent to which the investigative skills have developed. This measurement can only be achieved by assessing the answers given by the students (the product). The product would be then an automatic evaluation of the process because in all cases, the accuracy of the students' findings will affirm whether or not the skills of Noticing, Hypothesising and Experimenting have been applied and to what extent the skills have developed. In short, IOL tasks can be considered both process-and-product-oriented, where the value of the task lies in both.

IOL provides **authentic** data taken from corpora. Teachers can make use of the data to design activities for their student to practise applying their investigative skills.

In terms of **transfer of learning** outside of the classroom, IOL encourages reflective thinking in the sense that this approach tries to lead students to discover answers for themselves in the classroom in the expectation that they will be able to transfer the skills learnt outside of the classroom. Some of these skills will include knowledge of various types of collocations as well as colligational patterning.

Notes

1. CANCODE stands for 'Cambridge and Nottingham Corpus of Discourse in English'; the corpus was established at the Department of English Studies, University of Nottingham, UK, and is funded by Cambridge University Press (CUP). Sole copyright of the corpus resides with (CUP), from whom all permission to reproduce material must be obtained. The total corpus consists of five million words of transcribed conversations. The corpus tape-recordings were made in a variety of settings including private homes, shops, offices and other public places, and educational institutions, focussing on non-formal situations, across the islands of Britain and Ireland, with a wide demographic spread. For further details of the corpus and its construction, see McCarthy (1998).

References

Carter, R. (1998b). *Vocabulary: Applied Linguistics Perspectives, 2nd Edition.* London: Routledge

Channell, J. (1994). *Vague Language.* Oxford: Oxford University Press.

Collins Cobuild Grammar of English (1995) London: HarperCollins.

Finocchiaro, M. and C. Brumfit (1983). *The Functional-Notional Approach: From Theory to Practice.* Oxford: Oxford University Press.

Hunston, S., G. Francis and E. Manning (1997). Grammar and vocabulary: showing the connections. *English Language Teaching Journal* 51(3): 208-216.

Lewis, M. (1997). *Implementing the Lexical Approach.* Hove: Language Teaching Publications.

McCarthy, M. J. (1998). *Spoken Language and Applied Linguistics.* Cambridge: Cambridge University Press.

Quirk, R. and S. Greenbaum (1987). *A University Grammar of English.* Harlow: Longman.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985). *A Comprehensive Grammar of the English Language.* London: Longman.

Rastall, P. (1994). The prepositional flux. *IRAL* 32 (3): 229-231.

Sinclair, J. (1991a). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, J. (1996). The search for units of meaning. *Textus* IX: 75-106.

Stubbs, M. (1986). Lexical density: a technique and some findings. In M. Coulthard (ed.), *Talking about Text.* Birmingham: English Language Research, 27-42

Tognini-Bonelli, E. (1996). *Corpus Theory and Practice.* Italy: Tuscan Word Centre.