An English Spoken Academic Wordlist

Hilary Nesi
University of Warwick
CELTE
University of Warwick
CV4 7AL
UK
h.nesi@warwick.ac.uk

Abstract

This paper describes a project to develop an English spoken academic wordlist to complement the *Academic Word List* compiled by Coxhead in 1998. Syllabus designers consult learners' dictionaries when they are creating word lists for use in schools and colleges, and there is a fair degree of agreement at the most basic levels about which words are important and which words should be learnt. At more advanced levels, however, there is little interaction between syllabus designers and the designers of learners' dictionaries. A succession of university word lists, culminating in the *Academic Word List,* have been largely ignored by dictionary designers, despite the fact that most users of advanced learners' dictionaries are working with academic texts of one sort or another. This paper describes the process of creating a further academic word list, using data from the British Academic Spoken English (BASE) corpus, under development at the Universities of Warwick and Reading. It is argued that lists of this kind will be of value to learners studying in the medium of English, and to lexicographers producing reference materials for such learners.

Introduction

Wordlists are a crucial component of course and syllabus design for foreign language teaching the world over. In some educational systems they simply serve to remind materials writers and testers of the type of vocabulary learners will find most useful, but in other more regimented systems (particularly in the Far East) they can govern syllabus, textbook and test content to such an extent that at any given level only "wordlist words" are afforded lesson time [Tang & Nesi forthcoming].

The nature and content of these word lists is of great importance to the design of learners' dictionaries, because although learners may well look up words that they encounter outside the language lesson, those studying within a formal education system will pay most attention to wordlist words. A learners' dictionary can be seen to support the study process if the words on the wordlist are given special treatment, particularly regarding their productive use.

Specifications for the first few thousand words to be taught are much the same all over the world, with only minor variations to reflect local conditions. Beginners' wordlists have often been based on West's *General Service List of English Words* [1953], and nowadays also draw on recent corpus evidence, but they tend to be uncontroversial however they are compiled. Nation and Hwang [1995] report a large overlap between West's list and the high-frequency words in modern corpus counts, and Nation [2001:15] claims that there is about 80% agreement between any lists of high-frequency words drawn from well-designed corpora. Some advanced learners' dictionaries, of course, employ a restricted defining

vocabulary of the two thousand or so most frequent word families, roughly corresponding to the wordlist their users should have acquired in the first years of language learning. The practice of flagging high frequency words in these dictionaries has also helped syllabus designers to choose which words to teach. The wordlist for the 1999 revised College English Syllabus used throughout China, for example, draws (amongst other sources) on *Collins Cobuild* 1995 (words at level 5), *Longman Dictionary of Contemporary English* 1995 (the 3,000 most frequent words in speech and in writing), and *Oxford Word Power* 1993 (the 3,000 most frequent words).

Thus as far as wordlists for beginners are concerned, lexicographers and educationalists seem to be supporting each other's work. Both parties recognize the importance of the most frequent English words, essential for all learners because they supposedly account for about 90% of the running words in casual conversation and about 80% of the running words in written academic text [Nation 2001: 17].

Beyond this level, however, the choice of wordlist items, and their treatment in learners' dictionaries, becomes much more uncertain. Wordlist items for advanced study tend to have been selected by a mixture of methods, ranging from observation of learners' difficulties [Higgins 1967; Lynn 1973; Ghadessy 1979] to calculation of frequency in small corpora of university textbooks [Campion & Elley 1971; Praninskas 1972]. Until fairly recently the most widely discussed wordlist for tertiary level students was the *University Word List* [Xue & Nation 1984], developed by combining the lists compiled by Campion and Elley and Praninskas, and checking this new list against those of Lynn and Ghadessy. Although none of the major learners' dictionaries refer to it, the *University Word List* has been influential as a tool in English for Academic Purposes, serving as a syllabus component, as a yardstick by which to measure students' knowledge of the words they will need for academic study, and as a teaching tool (see, for example, the workbook "Check Your Vocabulary for Academic English" [Porter 2001]. Its successor is Coxhead's *Academic Word List* (AWL), which was developed from new corpus data in 1998, but it too has failed to attract much lexicographical attention as yet. (Reference is made to it in the study pages of the new *Macmillan English Dictionary,* [Rundell 2002] although entries for AWL words are not flagged.)

Range, as well as frequency, is an important element in the design of the *Academic Word List.* To be included word families have to occur at least 100 times in a 3,500,000 word academic corpus, and also at least ten times in each of the faculties represented by the corpus, and in more than half the 28 subject areas represented. Broadly speaking, the list is based on the principle that technical words have "a peak frequency of occurrence in one of several fields" [Yang 1986: 98] while academic words are context-independent, occurring with roughly equal frequency across disciplines [Cowan 1974]. Frequency counts alone are clearly not a very reliable measure of importance for all but the most common words, and must be qualified by evidence of distributional behaviour [Leech et al. 2001: 17]. Yet, apart from the traditional register labels and some distinction between spoken and written usage, English learners' dictionaries only present corpus data in terms of frequency, and do not systematically identify those words which are not extremely common, yet which have an important role to play across all academic disciplines.

Users of advanced learners' dictionaries need information about words that occur in academic contexts because many of them study in the medium of English, or encounter discourse similar to that found in university texts (factual writing, quality journalism, political debate, television and radio discussion). Increasingly, their advanced level English courses may also teach and test their knowledge of words on an academic wordlist. Coxhead's list is supported by corpus evidence, and is probably the best academic wordlist now available to compilers of advanced learners' dictionaries, as there is no separate and well-structured "academic" category in the larger corpora such as the British National Corpus. It is deficient, however, in that it is based entirely on data from textbooks – just one of several academic genres in widespread use. Other relevant genres include assignments, dissertations, and theses, which learners are likely to write rather than simply read, and lectures and seminars, in which learners may be involved both receptively and productively, as note-takers and as interlocutors.

The wordlist I am developing is designed to complement the *Academic Word List* by providing information about the use of words in spoken academic genres. It draws on the British Academic Spoken English (BASE) corpus, which is being compiled along similar lines to the Michigan Corpus of Academic Spoken English (MICASE), directed by John Swales and Sarah Briggs [2002]. The BASE corpus, which is freely available to researchers, currently consists of 127 hours of lecture and 32 hours of seminar recordings, of which 855,706 running words have already been transcribed. These represent about twenty-four and a half thousand different word forms; the type/token ratio in the corpus is 2.87.

In compiling the wordlist it has been necessary to take into account the fact that there are "degrees of technicalness" [Nation 2001: 198]. While highly technical words occur only very rarely outside their particular field, other technical terms are simply more common in one particular field than elsewhere, and may occur in quite a broad range of contexts. Likewise while some academic words express notions common to many disciplines rather than just one or two, possibly substituting for more frequent and informal near-synonyms, others will be almost empty of extra-linguistic reference, serving mainly rhetorical functions (the discourse-organizing words described as "procedural" by Widdowson [1983]). Thus compilation based on the principles of frequency and range will always require some dividing line to be rather arbitrarily drawn across the group of words which occur with medium frequency across a medium range of texts.

In view of the relatively small size of the corpus at present, the current academic word list consists of word families that occur more than three times in each of four broad subject areas, but which do not occur in Nation's list of the two thousand most frequent English words. Frequency measures will be increased, however, when the corpus reaches its target of two million words (slightly larger than MICASE). Division into broad subject groupings has proved problematic because the two models for the corpus, MICASE and the *Academic Word List*, both group disciplines according to administrative divisions in use at their respective universities. Thus MICASE follows the classification system used by the University of Michigan School of Graduate Studies, dividing the corpus into Humanities and Arts, Social Sciences and Education, Biological and Health Sciences, and Physical Sciences and Engineering, while Coxhead uses the four divisions of Arts, Commerce, Law and

Science, to match the faculties at the Victoria University of Wellington. The Universities of Warwick and Reading, where BASE is under development, have different and conflicting faculty systems, and also offer courses in less traditional disciplines. Warwick has three faculties (Arts, Social Studies and Science) while Reading has five (Letters and Social Sciences, Education and Community Studies, Agriculture and Food, Urban and Regional Studies, and Science). Clearly neither the MICASE nor the Victoria University of Wellington scheme would work well for the BASE corpus.

At present I am dividing the corpus according to the Kolb-Biglan classification of academic knowledge, as described by Becher [1989]. This incorporates the work of Biglan [1973], who created a discipline categorisation system based on questionnaire data from 222 American academics in 36 subject areas, and Kolb [1981], who used a psychometric test (the Kolb Learning Style Inventory) to test variations in learning style amongst 800 postgraduates with different disciplinary backgrounds. Becher drew on both sets of findings to create four categories of academic discipline: soft/pure (Languages, History etc), soft/applied (Education, Management, Law etc), hard/applied (Engineering, Medicine, Agriculture etc) and hard/pure (Chemistry, Mathematics, Computer Science etc.). This classification system is certainly not perfect, as it does not account for the many interdisciplinary courses which take place at Warwick and Reading, and the fact that within the same discipline different course modules often range in approach from hard to soft (as in Management for Engineers) or from applied to pure (as in Linguistics for English language teachers). Nevertheless it does have the very great advantage of being transferable across institutions, and it might therefore serve in the future as a common standard so that words from different academic corpora can be compared for frequency and range.

Many of the words in my current spoken academic wordlist match those in Coxhead's list. Some, however, do not, and reflect advances in technology since the AWL was compiled, or, more interestingly, generic differences in the data. BASE contains references to the *Internet* (18), *CD-ROMs* (16), the *web* (10), and, with great frequency, *video* (86) (doubtless prompted by the presence of recording cameras at the speech events). All of these are absent from the earlier lists. AWL contains *lecture*, marked as highly frequent, but not *seminar/seminars*, mentioned in the BASE corpus 170 times, or *handout/handouts* (mentioned 126 times). These words are relatively rare in written academic texts but are essential in academic speech communities which employ face to face teaching modes.

Some wide-range words in the BASE corpus which are not amongst Nation's two thousand most frequent words and are not listed in the *Academic Word List* reflect the interactive and interpersonal nature of spoken academic discourse. These include expressions of politeness such as *please*, which occurs 104 times, and *apology/apologize* which occurs 14 times ("If you speak Polish or have Polish ancestry, my apology if I have made a complete arse of how to pronounce this"). Vague words [Channell 1994] such as *stuff* (155) and *load/loads* (78) are frequent in the BASE corpus, but are not featured in the earlier lists. *Load* tends to be used to signify an approximate quantity, while *stuff* often serves a "placeholding" function in that it is used as a substitute in speech for a more precise term that a writer would have the time to recall. In the BASE corpus the social function of these vague words is also apparent, however, as they seem to be used to minimize threat and reinforce group solidarity ("You

just see like loads of Americans"; "you've got loads of information in the notes"; "God we've got a lot of stuff to cover I can see"; "a bit of fiddling about with the technical stuff").

Perhaps most noticeable, however, is the fact that academic speakers use words which express more explicit, extreme and subjective attitudes to their subject matter than those to be found in Coxhead's textbook corpus. In my data there are plenty of adjectival and adverbial markers of stance that are missing from the earlier lists. These include, for example, *boring / bored* ("this is probably a very boring part of the lecture") which occurs 29 times, and *bizarre* (10 occurrences) ("a really bizarre appointment"). Lecturers and seminar participants do not have to watch their words to the same extent as writers of published text, so there is less need to hedge opinions and more opportunity to vent personal feeling. The expression of stronger, more subjective views may also be an intentional device to generate enthusiasm for the topic, and indeed the majority of these evaluative or intensifying words in the corpus are positive in tone. Examples include *perfect* (89) ("I think the perfect example would be um visible light"), *highly* (85) ("a general but I think highly significant point"), *extremely* (83) ("these things are extremely effective"), *brilliant* (37) ("Carr performs a quite brilliant reconstruction of the role of ideology"), *amazing* (19) ("Anyone seen that film its amazing"), and *fantastic* (16) ("this is fantastic you're the next Keats"). *Fine*, (88) which does appear in Nation's 2000 word list in the sense of "thin" or "consisting of very small particles", occurs far more frequently in the BASE corpus as an expression of satisfaction or approval ("I think we will keep talking and that's fine"). In a manner typical of conversational as opposed to written academic style the adjectival forms sometimes occur as non-clausal fragments, as in these uses of *brilliant* - "you get nine points brilliant thanks very much indeed", *fine* - "OK you don't mind fine", and *perfect* -"alright perfect can you give me a name for your team".

Most of the new words in the Spoken Academic Word List are significantly more common in spoken than in written text. *Stuff,* for example occurs 274 times per million words in the spoken component of the British National Corpus, as opposed to 45 times per million words in the written component [Leech et al 2001]. Words in the BASE corpus which match those in Coxhead's Academic Word List, on the other hand, tend to be significantly more common in writing than in speech. *Political,* for example, which occurs 434 times in BASE and is included in the word family *policy* in the most frequent subsection of AWL, occurs 333 times per million words in the written component of the British National Corpus, but only 71 times per million words in the spoken component [Leech et al 2001]. Thus, as we might expect, the BASE corpus is found to contain both those features associated with published academic writing, as identified in earlier academic wordlists, and also those features associated with spontaneous spoken text, as recorded in non-academic spoken corpora.

Perhaps after all the identification of these words will result in only relatively minor additions to the current *Academic Word List,* which remains an excellent tool for the teaching and learning of English for Academic Purposes. Nevertheless a complete account of academic lexis requires authentic data from other than published written sources. As I think the few examples above have shown, the BASE corpus offers a window onto aspects of academic life that have hitherto remained largely unexplored by linguists and lexicographers, and for which many advanced learners need to be prepared.

Acknowledgements

The BASE corpus has been sponsored by the Universities of Warwick and Reading, The British Association for Lexturers in English for Academic Purposes (BALEAP), The British Academy, and (from April 2002) the UK Arts and Humanities Research Board.
The author particularly wishes to thank EURALEX for the Laurence Urdang Award which assisted the word list project described in this paper.

References

[Biglan 1973] Biglan, A., 1973. Relationships between subject matter characteristics and the structure and output of university departments, in: *Journal of Applied Psychology, 57 (3),* pp. 204-213, American Psychological Association, Washington, DC, US.

[Becher 1989] Becher, T., 1989. *Academic Tribes and Territories.* The Society for Research into Higher Education and Open University Press, Buckingham, UK.

[Campion & Elley 1971] Campion, M. E. & W. B. Elley, 1971. *An Academic Vocabulary List.* NZCER, Wellington.

[Channell 1994] Channell, J., 1994. *Vague Language.* Oxford University Press, Oxford.

[College English Syllabus Revision Team 1999] College English Syllabus Revision Team., 1999. College English Syllabus for students of Arts and Sciences - word list supplement (in Chinese). Beijing.

[Cowan 1974] Cowan, J., 1974. Lexical and syntactic research for the design of EFL reading materials, in: *TESOL Quarterly, 8 (4),* pp. 389-400, TESOL, Alexandria, Virginia, US.

[Coxhead 1998] Coxhead, A., 1998. *An Academic Word List.* English Language Institute Occasional Publication number 18. Victoria University of New Zealand.

[Coxhead 2000] Coxhead, A., 2000. *A new Academic Word List,* in: *TESOL Quarterly, 34 (2),* pp. 213-238, TESOL, Alexandria, Virginia, US.

[Coxhead & Nation 2001] Coxhead, A., & I.S.P. Nation 2001. *The specialised vocabulary of English for Academic Purposes,* in: J. Flowerdew & M. Peacock (eds.) *Research Perspectives on English for Academic Purposes.* Cambridge University Press, Cambridge.

[Ghadessy 1979] Ghadessy, M., 1979. Frequency counts, word lists, and materials preparation: a new approach, in: *English Teaching Forum,. 17 (1),* pp. 24-27, Washington, D.C., US.

[Higgins 1967] Higgins, J., 1967. Hard facts: notes on teaching English to Science students. Reprinted in J. Swales (ed.) *Episodes in ESP.* Pergamon Press, Oxford.

[Kolb 1981] Learning styles and disciplinary differences, in: A. Chickering (ed.) *The Modern American College.* Jossey Bass, San Francisco, US.

[Leech et al. 2001] Leech, G., P. Rayson & A. Wilson, 2001. *Word Frequencies in Written and Spoken English.* Longman, Harlow.

[Lynn 1973] Lynn, R., 1973. Preparing word lists: a suggested method, in: *RELC Journal, 4,* pp. 25-32, SEAMEO Regional Language Centre, Singapore.

[Nation 2001] Nation, I.S.P., 2001. *Learning Vocabulary in Another Language.* Cambridge University Press, Cambridge.

[Nation 1995] Nation, I.S.P. & K. Hwang, 1995. Where would general service vocabulary stop and special purposes vocabulary begin? in: *System, 23 (1),* pp. 35-41, Pergamon Press, Oxford.

[Paninskas 1972] Praninskas, J., 1972. *An American University Word List.* Longman, London.

[Porter 2001] Porter, D., 2001. *Check Your Vocabulary for Academic English.* Peter Collin Publishing, London.

[Rundell 2002] Rundell, M., 2002. (ed) *The Macmillan English Dictionary.* Macmillan, London.

[Swales & Briggs 2002] Swales, J., & S. Briggs, 2002. *The Michigan Corpus of Spoken Academic English (MICASE)* <http://www.lsa.umich.edu/eli/micase/micase.htm> accessed March 28 2002

[Tang & Nesi forthcoming] Tang, E., & H. Nesi, (forthcoming) Teaching vocabulary in two Chinese classrooms: intensive and extensive exposure in Hong Kong and Guangzhou, in: *Language Teaching Research.* Arnold, London.

[West 1953] West, M., 1953. *A General service List of English Words.* Longman, green and Co., London.

[Widdowson 1983] Widdowson, H., 1983. *Learning Purpose and Language Use.* Oxford University Press, Oxford.

[Xue & Nation 1984] Xue, G., & I.S.P. Nation, 1984. A University Word List., in: *Language Learning and Communication. 3 (2)*, pp. 215-229, John Wiley and Sons, London.

[Yang 1986] Yang, H., 1986. A new technique for identifying scientific/technical terms and describing science texts, in: *Literary and Linguistic Computing 1 (2)* pp. 92-103, Oxford University Press, Oxford.