Sarah Briggs & David Lee
ELI, University of Michigan

Developing a Lexical Database of Academic Spoken English (LDASE) for Language Testing: Problems & Prospects

Note: Having promised a SAWL (a word list) in our abstract, we are now aiming to deliver an LDASE (a lexical database)… read on for the reasons…

Goals

A frequency database of lexical items (words, multiwords, phrases) that:

1. is based on **academic spoken** English (the MICASE Corpus)
2. has **frequency & other statistical information for individual items,** plus various types of **distributional information** (e.g. which speech events, which disciplines/academic division, whether interactive or monologic)
3. **is accessible** on the web (restricted) in addition to paper versions (which are more like traditional lists, as opposed the database format we are proposing)
4. **flexible and customisable** (allows several views of the same database of words and frequencies: e.g. choose which columns of information to include or exclude; choose different cut-off frequencies; choose whether to group words by 'word family' or lemma or not at all)
5. **suitable for language test development purposes** (vocabulary test items, listening items, developing spoken prompts)

Problems with Present Word Lists / Issues to be Addressed

Issues	Existing Word Lists	Implications/Solutions
Based on **real speech events** in **academic** settings	No existing word list available based exclusively on spoken academic vocabulary. Lists for "General spoken English" exist (e.g. based on the British National Corpus), but not suitable for testing academic spoken vocab	LDASE will aim to fill this gap. The MICASE Corpus is 1.85 million words of speech collected from 15 different academic speech events, across all disciplines, within the University of Michigan during the period 1997-2001.
Frequencies of individual vocabulary items needed, along with other relevant information such as **range & dispersion** (a la Carroll *et al.* (1971) and James *et al.* (1994). Other word lists which give frequency information are Francis & Kučera (1964), for the Brown Corpus, and Johansson & Hofland (1989) for the LOB Corpus)	AWL (*Academic Word List*, Coxhead) for written academic English does not give such information for each individual vocabulary item.	LDASE will be implemented as **an on-line database** which allows users to select their own views of the data (e.g. view which academic disciplines a particular lexical item is most commonly used in; view type of speech event [lecture or dissertation defence or study group]) The idea is to allow flexibility: multiple views from one vocabulary database. The measures of **range** and **dispersion** help us go beyond **frequency** alone, which can be misleading: e.g. in the BNC, the scientific name *mucosa* (10 per million) is as frequent as *theirs* (10) or *shout* as a verb (10), while *magistrates* (21), *federation* (22), and *privatisation* (13) are all more frequent than *dirt* (10) and *arrow* (10)
No **GSL** ('General Service List') for '**general spoken language'**, apart from those based on frequency alone	West's (1953) GSL was used to filter Coxhead's results before the AWL could be determined. West's GSL for written language was intended as a measure of words most useful for 'general English', and was not **solely based** frequency	Should we have a Spoken GSL to filter LDASE through, in order to weed out '**non-academic**/ **general** spoken vocabulary'? Otherwise, how do we determine what is academic or not? **Options**: (1) Use *Keywords* analysis (Scott 1997, 2001) against the spoken part of the BNC (= British English)? (2) Use Peyawary's (1999) list of 'core interational English'? BUT: many seemingly 'ordinary' words in 'general English' may be used in special ways for special functions in **academic speech** → e.g. *way*
Multiple word units (MWUs): e.g. the conjunction *so that*, the preposition *in spite of*, and *at least* as an adverb), or semantic MWUs (e.g. *kick the bucket*). Spoken academic discourse may have its own distinctive MWUs	Not addressed in many word lists, including Coxhead's AWL. Addressed to some extent in Leech et al. (BNC Word Frequencies)	For language testing purposes, MWUs which function as a whole should be tested as single units (part-of-speech tags will reflect this as well)

Word families and **lemmas** (for 'word families' see Bauer & Nation 1993 and Nation & Waring 1997; for e.g. of lemmas in word lists, see the BNC Frequency book, Leech et al 2001)	'Word family' grouping: used in Coxhead's AWL, but no information on individual word forms: e.g. *concept* vs. *conception* vs. *conceptualize* vs. *conceptualization*. These different forms of the same word family may have very different distributions (e.g. *conception* has another sense too – see separate point below)	Database format of LDASE will allow viewing by lemma or by word family, but will also have word frequencies associated with individual word forms 'Lemma' grouping (= only inflectional (not derivational) affixes): used in the BNC Frequency book (individual word form frequencies also given)
Word senses versus word forms	Part-of-speech tags solve some of the problems, but still cannot distinguish a river bank from a financial bank, nor a coiled spring from a water spring or the season of spring.	LDASE will not address word senses to any large degree, apart from disambiguating items by POS. However, unlike other word lists (e.g. GSL, AWL) our source texts, the MICASE corpus, will be available on-line, so people can manually check for word senses if they so wish
Metaphorical and **idiomatic uses** of words: e.g. *a bear market*	Not addressed	Possibly address (manually)
Homography in untagged texts – e.g. May (month), may (verb); WHO (World Health Organisation), who (pronoun); Polish (from Poland), Polish (verb in initial position).	Addressed in some word lists	Addressed in LDASE (manually; however, word senses not distinguishable by POS or orthography cannot be addressed; e.g. *conception* [idea], *conception* [beginning], *conception* [pregnancy])
Part-of-speech (POS) information (grammatical word class)	Not addressed in some word lists (e.g. Coxhead's AWL).	LDASE will have POS information (e.g. noun/verb uses of the same word will have different entries & frequencies; e.g. *shout* (noun) = 3.76 per million words, *shout* (verb) = 1.78 per million)
Academic word lists should have **'general English' words interspersed** among the 'academic vocabulary' items, so that 'spoken academic words' can be seen in the context of a general language frequency list	Not addressed in Coxhead's AWL: only academic words are given	LDASE will allow viewing of academic vocabulary in the **context** of more general vocabulary. Users can choose to **show or hide** the "non-academic"/non-LDASE words

Productive frequency versus items actually problematic for learners	Not addressed in current word lists. For testing purposes, are the words derived from a 'production' corpus necessarily useful for language testing/scoring purposes?	What about other spoken academic words which are not necessarily **frequent** or **dispersed** enough but which learners have difficulty with? LDASE may include manual adjustments (?)
Counting word totals for spoken language (1): Interjections, discourse particles, hesitation markers, false starts/ truncated words, repetitions	Current word lists for 'general English' count all these as 'words' (thereby artificially inflating word totals)	LDASE will not count some of these (to be discussed!) towards word totals: this has possible implications for frequency values (e.g. false starts & truncations alone amount to about 3% of the MICASE corpus)
Counting word totals for spoken language (2): contractions and fused forms	Does not really affect written language word list	While everyone would agree that the contracted form *I've* and the fused form *gonna* both contain two morphemes each, a decision has to taken whether to count each form as constituting one word or two
Develop a spoken academic **phraseology** list?	Not addressed	Should we develop such a list, in addition to LDASE, where we list all commonly spoken 'academic phrases' (those which have particular academic discourse functions)? Goes beyond a 'word list', however, and it is not clear that an 'idiom' should be treated as a single lexical item rather than as a special combinatory use of several words
Productive vocabulary versus **Receptive** vocabulary	Not addressed.	Not addressed. Problem for spoken language tests: we want to test receptive spoken vocabulary, but filling in a blank or selecting a response from a multiple choice list is partly a productive task(?) Listening tests also test receptive phonological capacity in addition to vocabulary.

Contact Us

Sarah Briggs slbriggs @ umich.edu & David YW Lee dvdlee @ umich.edu
MICASE Project, English Language Institute, University of Michigan
TCF Building, 401 E. Liberty, Suite 350, Ann Arbor, Michigan 48104-2298, USA
Web Site: http://www.lsa.umich.edu/eli/micase/micase.htm

References & related works

Bauer, Laurie and I.S.P. Nation. 1993. Word families. *International Journal of Lexicography*, 6(3):1-27.

Carroll, John, Peter Davies & Barry Richman (1971) (eds.) *Word Frequency Book*. New York: American Heritage.

Coxhead, Averil and Nation, Paul (2001) The Specialised Vocabulary of English for Academic Purposes. In Flowerdew, J. and Peacock, M. *Research Perspectives on English for Academic Purposes*. Cambridge University Press: Cambridge.

Coxhead, Averil (2000) A New Academic Word List. *TESOL Quarterly*, 34(2): 213-238.

Francis, W.N. & Henry Kučera (1964) *Manual of Information to Accompany A Standard Corpus of Present-day Edited American English, for Use with Digital Computers*. Department of Linguistics, Brown University [orig. pub. 1964; rev. 1971, 1979, 1989]

Hofland, Knut & Stig Johansson (1982) *Word Frequencies in British and American English*. Longman: London.

Johansson, Stig & Knut Hofland (1989) *Frequency Analysis of English Vocabulary and Grammar: based on the LOB corpus*. Oxford: Clarendon Press.

James, Gregory, Robert Davison, Amos Cheung & Scott Deerwester (1994) *English in Computer Science: a corpus-based lexical analysis*. Hong Kong: Hong Kong University of Science and Technology & Longman Asia.

Johansson, Stig (1985) Word frequency and text type: Some observations based on the LOB corpus of British English texts. *Computers and the Humanities,* 19:23-36.

Leech, Geoffrey, Paul Rayson, & Andrew Wilson (2001) *Word Frequencies in Written and Spoken English: based on the British National Corpus*. Longman, London.

Peyawary, Ahmad S. 1999. *The Core Vocabulary of International English: a corpus approach*. Bergen: HIT-senterets publikasjonsserie Nr. 2/99.

Scott, Mike (1997). PC analysis of key words – and key key words. *System* 25 (2), Elsevier, pp. 233 – 245.

Scott, Mike (2001). Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith Tools suite of computer programs, in Ghadessy, M., Henry, A. and Roseberry, R. L. (eds.) *Small Corpus Studies and ELT: theory and practice*. John Benjamins, Amsterdam, pp. 47 – 67.

West, Michael (1953) *A General Service List of English Words with Semantic Frequencies and a Supplementary Word-list for the Writing of Popular Science and Technology*. London: Longman, Green & Co.

References on Vocabulary Assessment

Briggs, Sarah & Simpson, Rita. Using an academic corpus to evaluate the lexis of EAP tests. Language Testing Research Colloquium, St. Louis, MO.

Briggs, Sarah & Dobson, Barbara. Using a Spoken Language Corpus in the development of an EAP Listening Test. Poster presentation, Language Teaching Research Colloquium, Tokyo.

Read, John. 2000. *Assessing Vocabulary*. Cambridge: Cambridge University Press

Read, John and Carol A. Chapelle. 2001. A framework for second language vocabulary assessment. *Language Testing* 18 (1): 1-32

References on Multiwords/ Prefabs/Idioms

Aijmer, Karin. 1996. *Conversational Routines in English: convention and creativity*. London: Addison Wesley Longman.

Biber, Douglas and Susan Conrad. 1999. Lexical bundles in conversation and academic prose. In *Out of Corpora: studies in honour of Stig Johansson*, edited by Hilde Hasselgård and Signe Oksefjell, 181-190. Amsterdam: Rodopi.

Hudson, Jean. 1998. *Perspectives on Fixedness: applied and theoretical*. (Lund studies in English, 94). Lund, Sweden: Lund University Press.

Moon, Rosamund. 1998. *Fixed Expressions and Idioms in English: a corpus-based approach*. Oxford: Clarendon Press.

Nattinger, J. R. and J. S. DeCarrico. 1989. *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

Pawley, Andrew and Frances Hodgetts Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In *Language and Communication*, edited by J. C. Richards and R. W. Schmidt, 191-226. London: Longman.

Simpson, Rita & Dushyanthi Mendis. A corpus-based study of idioms in academic speech. Submitted to *TESOL Quaterly*.

Willis, Dave. 1990. *The Lexical Syllabus.* London: Collins ELT

References on MICASE and MICASE-based research

MICASE Web Site: http://www.lsa.umich.edu/eli/micase/micase.htm

Powell, Christina & Simpson, Rita. Collaboration between corpus linguists and digital librarians for the MICASE web search interface. In Simpson & Swales (eds.), 32-47.

Simpson, Rita C. & Swales, John M. (eds.) *Corpus Linguistics in North America: Selections from the 1999 symposium*. Ann Arbor: University of Michigan Press.

Simpson, Rita, Lucka, Bret & Ovens, Janine. Methodological challenges of planning a spoken corpus with pedagogical outcomes. Lou Burnard & Tony McEnery (eds.), *Rethinking Language Pedagogy from a Corpus Perspective: Papers from the third international conference on Teaching and Language Corpora*. Frankfurt: Peter Lang.

Swales, John M. & Burke, Amy. It's really fascinating work: Differences in evaluative adjectives across academic registers. The Third North American Symposium on Corpus Linguistics and Language Teaching, Boston, MA.

Swales, John M. & Malczewski, Bonnie. Discourse management and new episode flags in MICASE. In Simpson & Swales (eds.), 145-164.

General References on Vocabulary

Carter, Ronald. 1987. *Vocabulary.* Allen and Unwin: London.

Carter, Ronald and Michael McCarthy. 1988. *Vocabulary and Language Teaching*. Longman: London.

Coniam, David (in preparation) Word frequency and language proficiency.

Coniam, David (1995) Towards a common ability scale for Hong Kong English secondary school forms. *Language Testing* 12(2):184-95.

Schmitt, Norbert. 2000. *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, Norbert and Michael McCarthy, eds. 1997. *Vocabulary: description, acquisition and pedagogy*. Cambridge: Cambridge University Press.