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

<table>
<thead>
<tr>
<th>Issues</th>
<th>Existing Word Lists</th>
<th>Implications/Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Based on real speech events in academic settings</td>
<td>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</td>
<td>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.</td>
</tr>
<tr>
<td>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</td>
<td>Not addressed in many word lists, including Coxhead’s AWL. Addressed to some extent in Leech et al. (BNC Word Frequencies)</td>
<td></td>
</tr>
</tbody>
</table>

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 online 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 (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 international English’? BUT: many seemingly ‘ordinary’ words in ‘general English’ may be used in special ways for special functions in academic speech. e.g. way

- 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
- ‘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)
- ‘Lemma’ grouping (= only inflectional (not derivational) affixes): used in the BNC Frequency book (individual word form frequencies also given)
- Database format of LDASE will allow viewing by lemma or by word family, but will also have word frequencies associated with individual word forms

## 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 online, so people can manually check for word senses if they so wish

## Metaphorical and idiomatic uses of words:
- 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
- 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

## Counting word totals for spoken language (2):
- contractions and fused forms
- 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 & David Lee
ELI, University of Michigan

[Web Site: http://www.lsa.umich.edu/eli/micase/micase.htm]
References & related works

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.

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.

References on Multiwords/ Prefabs/Idioms

Simpson, Rita & Dushyantii 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

General References on Vocabulary

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