

**Spoken Lexicogrammar and
Discourse Patterns in the
Academy:
MICASE past, present and
future**

David Lee

English Language Institute (ELI)

University of Michigan

Outline

- **MICASE Past**
- **MICASE Present (MICASE-based research so far)**
- **MICASE Future (Phase 2 – Future developments)**

What is MICASE?

- Micase at a glance:
 - 1.7 million words of transcribed speech in 152 texts from about 200 hours of recordings of a variety of speech events within the university.
 - sampled “horizontally” across all the major academic divisions, and also “vertically”, moving from introductory undergraduate lectures to advanced graduate seminars
 - Speech events include lectures, seminars, student presentations, lab sections, office hours and study groups
 - native speakers of AmE (88% of tokens)

Why MICASE?

- EAP and ELI instruction:
- **Academic writing:** general characteristics and patterns
- **Discussion & Argumentation:** effective speaking in academic contexts – help students discover the appropriate lg for asking questions, stating a point of view, responding to opposing viewpoints, arguing a position.
- **Speaking fluency:** pronunciation and prosody

MICASE & small, specialised corpora

- large corpus → makes exploration and exploitation by potential end-users overwhelming and complicated
- may not have enough of specialized kinds of text (e.g. academic speech events such as office-hour consultations, dissertation defenses, student presentations in the classroom are not present in the BNC, but are in MICASE)

The Place of MICASE in the Corpus Universe

- go hunting for available corpora of American speech (<http://devoted.to/corpora>) → precious little available spoken American English (as opposed to **British** English).
- Santa Barbara Corpus of Spoken American English (CSAE); LGSWE, T2K-SWAL; Corpus of Spoken Professional English; LDC Switchboard & CallHome; Saarbruecken Corpus of Spoken English
- MICASE is currently the only publically available corpus of academic spoken American English, and fills a previous gap in the corpus universe

Academic writing and academic speech

- The study of academic **speech**, in comparison with academic writing -- much later start both for practical reasons and because of the perceived primacy of the written word in scholarly affairs
- one exception to this has been studies of introductory lecture discourse: primarily for the purpose of teaching lecture comprehension skills to ESL students (Flowerdew 1994; Tauroza 2001, for a recent review).

Summary of research findings to date, based on MICASE

- Anna Mauranen: **metadiscourse** (e.g., Mauranen 2001), **evaluation** (Mauranen 2000), and **hedging** (Lindemann and Mauranen 2001)
- John Swales: on the role of *point* and *thing* in metadiscoursal signaling (2001), on the role of multiple discourse markers such as *okay so now* (with Malczewski), on evaluative adjectives (Burke & Swales, in press), and on the limited utilization of sentence-initial ellipsis (e.g., *get my drift?*) (Swales 2002b).
- Swales (forthcoming) also provides provisional accounts of three MICASE research genres

Snippets of current research: some features of academic speech

- academic speech: really not hugely different from non-academic speech, although the differences and similarities are patterned and need to be studied
- textbooks based on the results of rather narrowly defined research, or the writer's intuitions about the kind of language that occurs in academic (or other) contexts → responsible for the discrepancy between the English that international students learn in their home country, and the English with which they are faced on arrival at a US academic institution

Non-standardised usages

- Working hypothesis: if features of 'language change' occur with some frequency in academic speech → changes within society as a whole are well on their way to becoming accepted and fixed

none

- *none* is followed by a plural verb **twice more often** than by a singular verb (e.g. *none of these are correct; none of the asperites are in a state of high stress*).

impact

- increasingly being used as a verb nowadays (1 out of 7 cases):
- interpersonal experiences in the family, um that may **impact** a current experience, it's really helpful to know. um
- i, like him or, don't like, u- even though of course, i- it'll **impact** how i write but um, i was trying to uh to get away...

data

- Swales & Feak (2000:111) report that "There is about a 2:1 preference for the uncountable" use of *data* in the MICASE corpus, where it is followed by a singular verb
- e.g. The data'**s** gonna be introspective / That data **is** still sitting there / They take the data, and they analyze **it** in a particular way

Lexicogrammatical Features of MICASE

- Modal Verbs
- Corpus-based studies of modal auxiliaries include Kennedy (2002), Mindt (1995) and, LGSWE (1999).

	Spoken UK English LLC (Coates 1983)	Written UK English LOB (Coates 1983)	Spoken & Written (Coates 1983)	BNC: All Spoken (Kennedy 2002)	Spoken AmE Academic MICASE
will	24.2	19.3	22.0	26.5 [1]	21 [3]
would	19.9	20.6	20.2	21.5 [3]	23 [2]
can	19.9	14.7	17.6	23.1 [2]	30 [1]
could	11.3	12.0	11.6	9.4	13.77 [4]
should	6.3	8.8	7.5	5.7	7.75
must	6.5	7.8	7.1	2.8	1.38
may	5.0	9.1	6.8	2.3	4.06
might	4.1	5.3	4.6	3.9	6.29
shall	2.8	2.4	2.6	1.3	0.16
Total	100	100	100	96.5	100

Table 1: Relative frequency of modals (Coates 1983) compared with MICASE frequencies (%)

Kennedy's figures here do not add up to 100% because I have excluded his percentages for *ought to* (0.6), *need to* (0.1), *dare* (0.1), and *used to* (2.8).

"General Eng" (Written & Spoken combined, mixed genres, Coates 1983) (same order also in LGSWE)	<i>will, would, can, could</i>
LLC (Spoken, Mixed)	<i>will, would, can, could</i>
BNC Spoken Mixed, BNC Conv, & LGSWE Spoken Conversation	<i>will, can, would, could</i>
MICASE (Academic speech, Mixed)	<i>can, would, will, could</i>
LGSWE Academic writing (Mixed)	<i>can, may, will, would</i>

Table 2: Ranked listing of the top four modals in several corpora, in descending order

MODAL FORM	PerM MICASE (AmE)	PerM ConvBNC (BrE)
can/ca n't/cannot	5,120	5,575
would/woulda/ 'd	4,010	3,736
will/wo n't/'ll	3,515	6,729
could/coulda	1,838	1,920
should/shoulda	1,034	1,045
might/mighta	839	856
may	542	151
must	184	716
shall	22	364

- *can* is significantly more common in British conversation than in American academic speech (along with *will*, *must*, and *shall*) and *would* and *may* are significantly more common in MICASE.

Artefact of the reference corpus we are comparing against?

MODAL FORM	PerM MICASE (AmE)	PerM SWB (AmE)
can/ca n't/cannot	5,120	3,607
would/woulda/ 'd	4,010	4,198
will/wo n't/'ll	3,515	2,420
could/coulda	1,838	1,505
should/shoulda	1,034	748
might/mighta	839	412
may	542	312
must	184	165
shall	22	19

- compared with the Switchboard corpus of prompted American telephone conversations, *would* is no longer significant, but *can* and almost all the other modals now seem to be significantly more frequent in MICASE



So which comparison gives a better picture of modal usage in MICASE?

- I really don't know
- Depends on your research question

Longman Grammar (p.491)

- *can* in academic **writing** commonly marks both ability and possibility
- no full, functional account of how this modal is frequently used in academic speech, as represented in MICASE

MICASE collocation/colligation

- 9,021 instances of *can/can't/cannot*, out of which about 39% are preceded by *you* (only 29% in the BNC Conversation corpus)
- interaction, and the achievement of particular interactional goals: collocates most frequently with 1st and 2nd person pronouns (*you can/can't, we can, I can, can you*)
- demonstrating or teaching how to do something; suggesting or even directing (in a very non-threatening and not-so-imperative way)

- ... and then sloughs are over and up high maybe *you can* see that those are areas that are very they 're ...[a]
- ... and what 's this is the melting point so *you can* either melt it try to dilute it solid put it ...[p]
- ... treat this as a lump sum heat capacity where *you can* uh treat the temperature inside it as the exact...[p]
- ... person him or herself whatever that means and then *you can* uh think of the saint also as a creator of ...[p]
- ... just like there 's different levels of organization *you can* look at in humans *you can* look at the um *you can* look at the cellular level *you can* look at ...[p]
- ... so you know what the elevations are using those *you can* calculate the pressure *you can* either do that or before you even start doing that *you can* look at this ...[p]

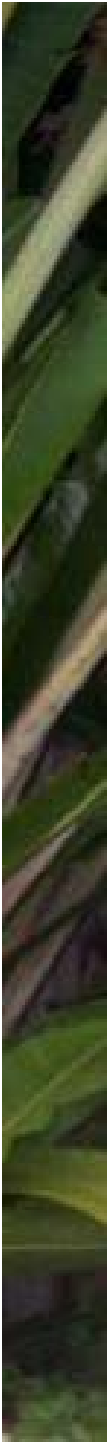
one grammar reference book...

- “The *can* of permission is sometimes used as an indirect way of making a strong suggestion or giving a command: "Don't be afraid. You *can* tell your critics that I am on your side." Instead of speaking indirectly, the speaker can achieve much the same effect by using direct commands... *can* in these sentences is effectively used to soften the tone.”

Suzuki, p.287

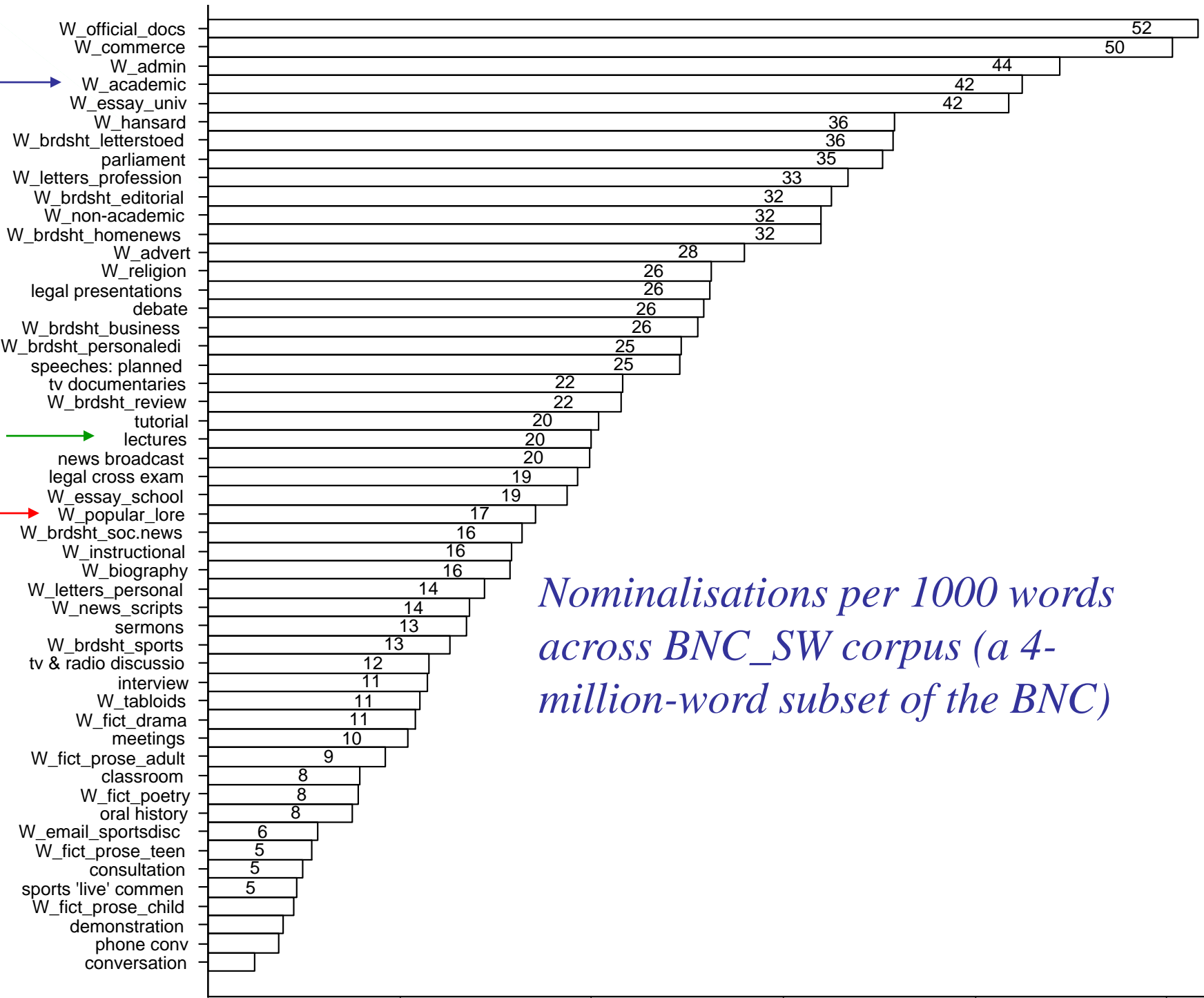
"polite directing/advising" function

- the 'office hour' consultation
- *you can* : functions to give advice/reassurance with technical problems or to give explanations relating to them
- ... you **can** still call it a T-statistic if your sample is large enough [p]
- ...you **can** analyze both of them but one at a time [p]
- ... if it's a pair design you **can** enter the differences [p]
- ... you **can** have the list, do your X-bar [p]
- ... you **can** put in L-one just to show what that looks like [p]
- ... if you want it on L-one, you **can** just hit enter [p]
- ... so you **can** call it a Z if you wish [p]

- 
- The *you can* lies somewhere in between the *one can* of formal discourse, and the *we can* of closer affiliation. So the *you cans* seem then to mean that the students CAN do these things without the help of the ‘experts’

Nominalisations

- *TION_N*/*MENT_N*/*ITY_N*/*NESS_N*
- occur about 17 times per thousand words in MICASE



*Nominalisations per 1000 words
across BNC_SW corpus (a 4-
million-word subset of the BNC)*

Nominalisations (cont'd)

- Closest to W_pop_lore in the BNC (e.g. magazines such as *Gardeners' World*, *Country Living*, *She*, *Yorkshire Life*, *Punch*, *The Face*, *Esquire*)
- not too different from British university lectures in terms of how heavily nominalised it is
- suggests that if NNS find academic discourse difficult to follow, the problem could lie in the kinds of nominalisations they encounter rather than the frequency of them or perhaps nominalisations are not a problem at all

MICASE Phase II Projects

- POS-tag; lemmatise; XMLise; segment texts into C-units(communucative units); annotate files with discourse/pragmatic codes; synchronize the text transcripts with the audio files; create a lexical frequency database of academic speech; develop a new web search interface (with X-Sara as the server); conduct research for a planned book on the grammar of academic speech.

POS Taggers, Spoken Language and Dialects: unlikely bedfellows?

- American English spoken transcription conventions and American lexical items – problem for CLAWS
- sometimes intervening adverbs or interjections (e.g. *um*, *erm*) cause errors: e.g. *_VM *_UH *_V?0
- some of these are also issues for end-users who want to retrieve items (you've got to know the conventions)

MICASE POS-tagged

MICASE Corpus text (file adv700ju023, tagged using CLAWS C7 tag set)

```
<U WHO="S1"> so_RR <PAUSE DUR=":01" TYPE="FINAL"> i_PPIS1  
see_VV0 that_CST you_PPY 're_VBR from_II Hartland_NP1  
Michigan_NP1 <U WHO="S2"> yes_UH </U> this_DD1 is_VBZ <PAUSE  
DUR=":01" TYPE="CONT"> right_JJ up_II the_AT road_NN1 </U>  
<U WHO="S2"> mhm_UH <PAUSE DUR=":01" TYPE="CONT"> like_RR  
forty_MC minutes_NNT2 from_II here_RL <U WHO="S1"> yeah_UH </U>  
mhm_UH </U>  
<U WHO="S1"> okay_RR <PAUSE DUR=":01" TYPE="CONT"> and_CC  
uh_UH <PAUSE DUR=":01" TYPE="CONT"> you_PPY say_VV0 that_CST  
you_PPY 're_VBR interested_JJ in_II prebusiness_NN1 and_CC  
economics_NN1 </U>  
<U WHO="S2"> i_PPIS1 was_VBDZ i_PPIS1 do_VD0 n't_XX think_VVI  
that_CST i_PPIS1 am_VBM anymore_RR  
<EVENT DESC="LAUGH"> </U>
```

XMLisation & audio-linking

- XMLise MICASE texts in preparation for audio-transcript synchronisation and the new web search engine (“X-Sara”)
- employ the latest multimedia standards on the web
- partnering with a team of web-based multimedia experts at Michigan State University

- SMIL (Synchronized Multimedia Integration Language)
- time-synchronized SMIL presentation with an audio timeline and an accompanying scrolling text display in a RealPlayer window embedded into a custom web browser window”
- with audio links, users can go from any concordanced line to its associated sound recording

Segmentation

- C-unit = spoken equivalent to the ‘sentence’ in written language, “a unit with optimal syntactic independence, in that it is not part of a larger syntactic unit, except by means of coordination” (Leech 1999:108). Thus, syntactically free-standing units such as *indeed*, *sorry*, *Jimmy*, or *no problem*, which have no finite verb, are each a C-unit.
- “any adjective followed by any noun” → × instances where a speaker ends with an adjective and begins a new syntactic unit with a noun (e.g. *and in effect that's true < > people study heterochrony today, and think about these...* [LEL115SU107]).

Pragmatic and discourse-level coding in the file headers and/or within the texts

- analyze and teach—using authentic data from the corpus—functions such as **soliciting and offering advice, giving instructions, expressing disagreement, or asking questions**
- retrievable in conjunction with the various speaker and speech event categories already encoded: e.g. investigate differences in the use of question types or advice-giving by professors in different speech events (e.g. office hours versus lab sections).

BNCweb Demo (a foretaste of things to come: MICASWeb)

- The BNC
 - The BNC World Edition is a 100-million-word corpus (more like 97.6 million) of spoken and written British English, of which roughly 10 million is transcribed speech (both spontaneous conversations and task-orientated speech such as lectures, committee meetings, sermons, interviews and various TV and radio broadcasts).

- **BNCweb**: a web-based client program for searching and retrieving lexical, grammatical and textual data from BNC; a research and teaching web-based environment: browsing, concordancing, collocation searching & calculating, distributional analyses, and much more – all within an intuitive, user-friendly interface.
- With BNCweb (and our future MICASEweb), users can easily create subcorpora and restrict their searches to that specific subset of the larger corpus, perform various functions on the data, retrieve a list of collocations, and go directly from a concordance line to the associated sound clip.

Distribution Function

- [DEMO: Concordance of *lovely* and its distribution]

All this and more will be yours...

- everything demonstrated today will also be possible on the future MICASEweb (+ more).
- When completed, MICASEweb will provide the first publicly accessible corpus of spoken English with text transcripts linked to the audio recordings for easy access, and all within a sophisticated concordancing package.

Fin

完

~~~~ \* ~~~~