MICASE

MICASE Manual

The Michigan Corpus of Academic Spoken English

Version 1, dated July 22, 2003

www.lsa.umich.edu/eli/micase/micase.htm

The English Language Institute
University of Michigan

NOTE for those reading this PDF file on screen:

All document structure numbers in this document (i.e., cross-references to section headings or tables) are clickable hyperlinks (not visibly marked) which you may use to jump directly to the relevant section. All headings in the table of contents and all web addresses (URLs) and e-mail addresses are also 'live' hyperlinks.

Questions or comments about any aspect of MICASE may be sent to micase@umich.edu or to MICASE, English Language Institute, University of Michigan, TCF Building, 401 E. Liberty, Suite 350, Ann Arbor, Michigan, MI 48104-2298, USA.

Contents

Chapter 1

Introduction and history

In late 1997, the English Language Institute (ELI) at the University of Michigan started a major research project to create a resource for studying academic speech. The goal of the first phase of the project was to record and transcribe close to 200 hours (approximately 1.7 million words) of academic speech from across the university. In June 2001, we finished the recording goal, with over 190 total hours recorded. In April 2002, we completed transcribing and proofreading all the transcripts. (The digital sound recordings were transcribed with the help of a computer program called SoundScriber, developed by former research assistant Eric Breck.)

Analysis is supported by a customized search engine developed for us by the Humanities Text Initiative section of the University Library. The entire corpus is now available on-line at www.hti.umich.edu/m/micase. This search engine is notable for the large number of speaker and speech-event categories that can be selected, and the user-friendly interface.

The MICASE corpus is a spoken language corpus of approximately 1.7 million words (nearly 200 hours) focusing on contemporary university speech within the microcosm of the University of Michigan, in Ann Arbor, Michigan. This is a typical large public research university with about 37,000 students, approximately one-third of whom are graduate students. Speakers represented in the corpus include faculty, staff, and all levels of students, and native, near-native and non-native speakers.

The ELI has committed resources to MICASE for a series of interlocking reasons. First, before this project, there was no database of this kind available. Second, we strongly suspect that once we are able to examine the corpus for recurrent grammatical and phraseological patterns, we will find many divergences from those described in current grammar and vocabulary books, which have largely relied on introspection or on features of written texts. MICASE thus provides authentic material in sufficient quantity to redefine our concepts of academic speech. Third, we eventually hope to be able to track generalized changes in speech patterns as people gain experience of university culture. (Although we know quite a lot about how academic writing evolves as students progress, our current perceptions of speech changes within academic cultures are

largely anecdotal.) Fourth, with all this new information, we—and others elsewhere—will be in a better position to develop more appropriate ESL and English for Academic Purposes teaching and testing materials, and to evaluate how best to incorporate corpus work into EAP programs.

The project is managed by Dr Rita Simpson, with Professor John Swales and Dr Sarah Briggs acting as faculty and testing advisors, respectively, and Dr. David Lee as a post-doctoral research fellow. Current research assistants are Susan Hoyenga, Ruth Kraut and Sheryl Leicher. Professor Anna Mauranen of Tampere University, Finland served as an external consultant from 1998 to 2001.

Chapter 2

Corpus content and structure

In this chapter, basic information is given about the types of texts you can find in MICASE, allowing an overall view of the types of speakers and speech events that have been included.

2.1 Academic speech

Academic events vary widely in their tone, substance, and length. The MICASE corpus includes speech events that range in length from 19 to 178 minutes, with word counts ranging from 2805 words to 30,328 words. In the MICASE corpus, *academic speech* is defined as that speech which occurs in academic settings. In other words, it is not predefined as something like "scholarly discussion." In academic settings, we might, for example, find such speech acts as jokes, confessions, and personal anecdotes, as well as definitions, explanations and intellectual justifications. Therefore, the MICASE researchers have taken pains to record a wide variety of academic speech events. Most speech events are fully recorded, from beginning to end, because the beginnings and ends of academic speech events may be of particular interest to researchers.

Certain events that occur on campus may not qualify as academic speech events, either because they would not be significantly different if they had occurred in other locations, or because they are not particular to a university community's educational mission or research agenda. For example, we did not record food-ordering sequences in university food outlets or office talk among co-workers in various university support staff departments and offices.

The 152 speech events in the corpus include:

- small and large lectures (62)
- public interdisciplinary or departmental colloquia (13)
- student presentations (11)
- discussion sections (9)
- seminars (8)

undergraduate lab sessions (8)
office hours (8)
study groups (8)
lab group and other meetings (6)
advising consultations (5)
dissertation defenses (4)
one-on-one tutorials (3)
interviews (3)
campus/museum tours (2)
service encounters (2)

2.2 Structure of the corpus

The corpus was designed to be balanced, as much as possible, across several categories of academic speech events as well as across the major academic divisions within the university. Academic events in the professional schools (i.e., medical, dental, business, law) were excluded. The range of speech events includes monologic and interactive speech; undergraduate and graduate students; junior faculty, senior faculty, and staff; and native, near-native, and non-native speakers of English. Furthermore, an attempt was made to get approximately equal amounts of speech from male and female speakers within each academic division. For a detailed breakdown of the word counts and percentages of speech by each category of speaker and within the two major speech event categories, see Chapter 4 of this manual.

Each speech event in MICASE is categorized according to various contextual attributes, and these attributes can be found in the header of each transcript. Speech event attributes include the type of event, the subject area of the event, the extent to which an event is monologic or interactive, as well as the academic role or level of the majority of participants (e.g., whether the class was a graduate or an undergraduate class, or whether a meeting was primarily of senior faculty members). A description of all the speech event attributes and their corresponding codes is found in the table in section 2.3. Similarly, all speakers in the corpus were classified according to several different demographic variables (e.g., gender and age), which can be found in the table in section 2.4.

2.3 Speech event attributes

SPEECH EVENT TYPE		
CLASSROOM EVENTS *NOTE: All classroom speech events are defined externally by the university regardless of the actual speech event characteristics, except in cases where prepared student presentations constitute the majority of the speech, in which case the event type is Student Presentations.*		
SMALL LECTURES	LES	Lecture class; class size = 40 or fewer students
LARGE LECTURES	LEL	Lecture class; class size = more than 40 students

DISCUSSION SECTIONS	DIS	Additional section of a lecture class designed for maximum student participation; may also be called recitation
LAB SECTIONS	LAB	Lab sections of science and engineering classes; may include problem solving sessions
SEMINARS	SEM	Any class defined as a seminar (primarily graduate level)
STUDENT PRESENTATIONS	STP	Class other than a seminar in which one or more students speak in front of the class or lead discussion
NON-CLASS EVENTS		
ADVISING SESSIONS	ADV	Interactions between students and academic advisors
COLLOQUIA	COL	Departmental or University-wide lectures, panel discussions, workshops, brown bag lunch talks, etc.
DISSERTATION DEFENSES	DEF	Ph.D. theses defenses
INTERVIEWS	INT	Interviews for research purposes
MEETINGS	MTG	Faculty, staff, student government, research group meetings, not including study group meetings
OFFICE HOURS	OFC	Held by faculty or graduate student instructors in connection with a specific class or project
SERVICE ENCOUNTERS	SVC	Library, computer center, financial aid office services
STUDY GROUPS	SGR	Informal student-led study groups, one time or on-going
TOURS	TOU	Campus, library, or museum tours
TUTORIALS	TUT	One-on-one discussions between a student and an instructor or peer tutor

ACADEMIC DIVISION AND DISCIPLINE	
Academic discipline corresponds to individual university departments when applicable; otherwise assigned as miscellaneous. *Academic division* refers to one of four divisions defined according to the Horace H. Rackham School of Graduate Studies classification of departments.	
BIOLOGICAL AND HEALTH SCIENCES (BS)	Includes Biology, Biochemistry, Dentistry, Genetics, Immunology, Natural Resources, Neuroscience, Nursing, Pathology, Pharmacy, Physiology, Public Health
PHYSICAL SCIENCES AND ENGINEERING (PS)	Includes Astronomy, Chemistry, Computer Science, Engineering (all), Geology, Mathematics, Physics, Statistics, Technical Communication
SOCIAL SCIENCES AND EDUCATION (SS)	Includes Anthropology, Business Administration, Communication, Economics, Education, History, Public Policy, Political Science, Psychology, Social Work, Sociology, Urban and Regional Planning
HUMANITIES AND ARTS (HA)	Includes Area Studies (all), Architecture, Classics, Comparative Literature, English, Fine Arts (all), Foreign Languages, History of Art, Information and Library Science, Linguistics, Philosophy, Women's Studies

PARTICIPANT LEVEL		
Corresponds to the level of the majority of students for classes, or participants for other events.		
JUNIOR UNDERGRAD	JU	First and second year undergraduates
SENIOR UNDERGRAD	SU	Third year and above undergraduates
MIXED UNDERGRAD	MU	Mixed undergraduates
JUNIOR GRADUATE	JG	First and second year or Master's level graduate students
SENIOR GRADUATE	SG	Third year and above Ph.D. students
MIXED GRADUATE	MG	Mixed grad students
JUNIOR FACULTY	JF	Lecturers and Assistant Professors
SENIOR FACULTY	SF	Associate Professors and above
MIXED FACULTY	MF	Mixed faculty
RESEARCHER	RE	Non-teaching researchers
POST-DOC FELLOW	PD	Post-doctoral research fellows
STAFF	ST	Non-teaching University employees
VISITOR/OTHER	VO	Non-UM or non-academic affiliates
MIXED	MX	Mixed faculty, staff, students
PRIMARY DISCOURSE MODE		
Refers to the predominant type of discourse characterizing the speech event.		
MONOLOGIC	MLG	One speaker monopolizes the floor, sometimes followed by question and answer period
PANEL	PNL	Several consecutive monologues usually followed by multi-speaker interactions
INTERACTIVE	INT	Interactional discourse involving two or more speakers
MIXED	MIX	No one discourse mode is predominant

2.4 Speaker attributes

CATEGORY	CODE	DEFINITION/COMMENTS
GENDER		
FEMALE	F	
MALE	M	
AGE GROUP		
17 - 23	1	
24 - 30	2	
31 - 50	3	
51 and older	4	
ACADEMIC ROLE		
JUNIOR UNDERGRAD	JU	First and second year undergraduates
SENIOR UNDERGRAD	SU	Third year and above undergraduates
JUNIOR GRADUATE	JG	First and second year or Master's level graduate students
SENIOR GRADUATE	SG	Third year and above Ph.D. students
JUNIOR FACULTY	JF	Lecturers and Assistant Professors
SENIOR FACULTY	SF	Associate Professors and above
RESEARCHER	RE	Non-teaching researchers
POST-DOC FELLOW	PD	Post-doctoral research fellows
STAFF	ST	Non-teaching University employees
VISITOR/OTHER	VO	Non-University of Michigan affiliates
NATIVE SPEAKER STATUS		
NATIVE SPEAKER	NS	Native speakers of North American English
NATIVE SPEAKER OTHER	NSO	Native speakers of non-American English
NEAR NATIVE SPEAKER	NRN	Non-native speakers who consider English as their current dominant language and who appear to have native-like fluency and grammatical proficiency.
NON-NATIVE SPEAKER	NNS	Non-native speaker of English other than near-native speakers
FIRST LANGUAGE		
Only shown when first language is other than North American English.		

Chapter 3

Data collection, transcription, and mark-up

This chapter describes how the corpus data were recorded, transcribed and marked up. It gives a full account of how words are spelled in the (various versions of the) corpus and other aspects of the transcription system

3.1 Data collection and recording methodology

Because MICASE aimed to record a wide range of academic speech, our sampling goals spanned fifteen different types of speech events and four major academic divisions within those types (e.g., humanities, social sciences, physical sciences and engineering, and biological sciences). (See section 2.3 above.) We adopted stratified random sampling as our preferred method of sampling. Each recording is classified according to speech event type, a pre-assigned number indicating the academic discipline, two letters representing the majority of participants in the event (e.g. junior undergraduate, senior graduate, senior faculty), and a final three digit sequence to track chronologically when the tape was recorded. For example, transcript number LEL115SU015 is a recording of a large lecture (LEL) in anthropology (115), at the senior undergraduate level (SU), and is the 15th speech event recorded for MICASE.

All recordings were made with a digital audio tape recorder with two external stereo microphones, and at selected events, a video recorder. Two researchers attended most speech events in order to identify speakers and facilitate transcription by taking field notes about nonverbal contextual information; however, in small groups (e.g. advising sessions, office hours, study groups) where an observer's presence would have been intrusive, the research assistants left the room after the equipment was set up. All speech was recorded with written consent from the major speakers and verbal consent from other participants. Demographic information (gender, age group, university position, and native language) was collected from each speaker on a form distributed at the end of each event. The speaker information is included in the header of each transcript

and is also entered into a separate database. All DAT recordings are captured and stored as MP3 format sound files for use with our computer transcription program, SoundScriber.

3.2 Corpus mark-up and annotation

The construction of MICASE was based on guidelines established by the Text Encoding Initiative (TEI) and files were originally marked up in SGML. We haved subsequently updated the corpus DTD (Document Type Definition) and converted all the files to the XML format. This was in order to realize our corpus development plans, which include a new search interface and the streaming web delivery of the sound recordings, synchronized with the transcripts.

At present, only the orthographically transcribed version of the corpus is available. Future releases of MICASE, however, will have various kinds of linguistic annotation added: parts of speech, lemmas, and discourse-pragmatic categories. The part-of-speech tagger we will use on MICASE is the CLAWS tagger developed at Lancaster University, UK (Appendix B gives the tag set adapted for use with our corpus). In addition, manual and semi-automatic (rule-based) corrections and refinements to the CLAWS tagging will be undertaken. Details will be given in the next update of this manual, along with the tagged corpus.

3.2.1 Spelling, transcription and mark-up conventions

The MICASE orthographic transcription conventions and mark-up system are intended to allow for ease of readability, while including enough detail to ensure adequate comprehension from the text of the transcript alone. To this end, we use standard orthography in the case of most words, except for select situations where standard conventions may cause confusion, and for a limited number of lexicalized abbreviations and grammatical constructions (e.g., *cuz, gonna, hafta, sorta*, and several others). We do not use standard punctuation, but instead mark pauses of varying lengths with commas, periods, and ellipses. We also use question marks to identify phrases that function pragmatically as questions.

All backchannel cues and hesitation or filler words were transcribed using a set number of normalized orthographic representations that disregard minor phonetic variations. These, like overlaps and interruptions, are transcribed in a way that illustrates their sequential occurrence, but still indicates which speaker holds the floor.

We originally used a customized set of SGML tags adapted from the TEI conventions, which have since been converted to XML. Additionally, all the speaker demographic information and recording information is tagged in the header. Our transcripts were first created using Author Editor, an SGML text editing program. After the XML conversion, we used XMLSpy.

The tables below give a complete description of the spelling, transcription, and mark-up conventions (they are also available on the transcription conventions page on the MICASE website).

Table 3-1 Spelling Conventions

	RULE or GUIDELINE	EXAMPLES
GENERAL	Standard orthography is used for most words, even though they may not be fully pronounced, may be pronounced with a foreign accent, etc. In general, phonologically reduced forms are not represented, except as noted below.	
CAPITALIZATION	Only proper nouns (names, departments, course titles, organizations, etc.) are capitalized (in addition to acronyms; see below). Neither the beginnings of turns nor the pronoun 'i' are capitalized.	Dr Hales received his M-S and B-S degrees at Stanford in nineteen eighty-two. his PhD at Princeton in eighty-six under the Harold W Dodds Honorific Fellowship... oh, i i think i know what you're getting to.
FILLED PAUSES, BACKCHANNEL CUES, EXCLAMATIONS, etc.	All hesitation and filler words, backchannel cues, and transcribable exclamations are spelled out, as shown on the right.	**Hesitation/Filler Words/Backchannels**: hm, hm', huh, mm, mhm, uh, um, mkay **Yes/No Responses**: yes: mhm, mm, okey-doke, okey-dokey, uhuh, yeah, yep, yuhuh no: uh'uh, huh'uh, 'm'm, huh'uh **Exclamations/Doubt/Misc**.: ach, ah, ahah, gee, jeez, oh, ooh, oop, oops, tch, ugh, uh'oh, whoa, yay

CONTRACTIONS and LEXICALIZED REDUCED FORMS	All standard contractions of *is, am, are, had, have, would, not* are represented, including [noun + *has been/have been/is*].	i'd, i've, i'm, i'll, she's, she'll, he's, they've, etc. that'll, it'll, there're etc.
	Different forms of modals + have are represented	coulda, could've, couldn't, couldn've, couldna, woulda, would've, wouldn't, wouldn've, wouldna, shoulda, should've, shouldn't, shouldn've, shouldna
	Lexicalized phonological reductions are limited to those listed on the right.	betcha, cuz, 'em (=them), gimme, gotta, hafta, kinda, lookit (as vocative only), lotsa, lotta, oughta, sorta, wanna
ACRONYMS, ABBREVIATIONS, LETTERS AS VARIABLES	Acronyms are written in all caps. Three commonly abbreviated titles are left as abbreviations, but without periods.	Exception: PhD (no hyphens, no period) Dr, Mr, Mrs (not spelled out)
	An acronym pronounced as a word is run together as one word.	NASA, TOEFL
	[On-line version only] When an acronym is spelled out, it appears in all caps with hyphens between each letter (except PhD). [In the CD-ROM/ distributed version and the tagged version, these hyphens have been removed.]	C-I-A F-B-I E-L-I L-S-and-A
	Letters used as variables in math and science are written in all caps with hyphens between modifying or adjoining elements.	X-Y axis N-squared, X-to-the-N-minus-one
HYPHENS	Standard hyphenation rules apply, as in the Chicago Manual of Style, where they exist.	pre-med, pre-calc, pre-law, mid-thirties mid-nineteen-ninety-nine, pre-Christian, non-Euclidean, non-native
NUMBERS	All numbers are fully spelled out as words. Standard hyphenation rules apply, with some additional guidelines: page numbers, course numbers, and room numbers are all hyphenated.	nineteen ten nineteen twenty-nine page one-fifty-seven Poli Sci one-sixty room thirty-twelve

REPETITIONS and REPAIRS	All repetitions of a word, partial word or phrase are transcribed. Truncated or cut-off words have a hyphen at the end of the last audible sound/letter. An underscore at the end of a word indicates a false start in which a whole word is spoken but then the speaker re-starts the phrase.	it's no longer than a than a, calendar year... so, come on up, grab yourself a ins- implement of destruction. well, it will be_ it's sort of_ it's a management human-resource kind of job...
FOREIGN WORDS	Foreign words are spelled as in the original language when it uses a roman alphabet; otherwise, an approximate phonetic transliteration is used.	and see what, the Buddha, was s- was saying um, the *tatha, gata*, Sanskrit's a really interesting language...
PRONUNCIATION VARIATIONS	As mentioned above, minor pronunciation variations are not represented in the spelling, with the exception of the contractions and lexicalized forms listed in this table.	

Table 3-2 Transcription and Mark-Up Conventions

XML TAG or SYMBOL	MEANING/DESCRIPTION	APPEARANCE IN ONLINE TRANSCRIPTS (HTML VERSION)
SPEAKER ID		
<U WHO=S1>, <U WHO=S2>, etc.	Speaker IDs, assigned in the order they first speak.	**S1:** at the beginning of each turn or interruption/backchannel.
<U WHO=SU>, <U WHO=SU-f>, <U WHO=SU-m>	Unknown speaker, without and with gender identified	**SU:** **SU-f, SU-m**
<U WHO=SU-1>	Probable but not definite identity of speaker	**SU-1:**
<SS>	Two or more speakers, in unison (used mostly for laughter)	**SS:**
PAUSES		
<PAUSE DUR=:05>	Pauses of 4 seconds or longer are timed to the nearest second.	<P: 05>

, [for untagged and on-line versions. Tagged version: <PAUSE DUR=":01" TYPE="CONT">]	Comma indicates a brief (1-2 second) mid-utterance pause with non-phrase-final intonation contour.	,
. [for untagged and on-line versions. Tagged version: <PAUSE DUR=":03" TYPE="FINAL">]	Period indicates a brief pause accompanied by an utterance final (falling) intonation contour; not used in a syntactic sense to indicate complete sentences.	.
... [for untagged and on-line versions. Tagged version: <PAUSE DUR=":03">]	Ellipses indicate a pause of 2-3 seconds	...
OVERLAPS		
<OVERLAP>...</OVERLAP>	This tag encloses speech that is spoken simultaneously, either at the ends and beginnings of turns, or as interruptions or backchannel cues in the middle of one speaker's turn. All overlaps are approximate and shown to the nearest word; a word is generally not split by an overlap tag.	Text of overlapping speech is in blue.
BACKCHANNEL CUES and FAILED INTERRUPTIONS		
Embedded utterance (<U> tag within a <U> tag)	Backchannel cues from a speaker who doesn't hold the floor and unsuccessful attempts to take the floor are embedded within the current speaker's turn, and not shown as a separate line/paragraph.	[S3: Text of embedded speech is in orange and surrounded by orange square brackets.]
Embedded and overlapped utterance (<OVERLAP> tag within an embedded utterance)	Backchannel cues or unsuccessful interruptions that overlap with the main speaker's speech.	[S3: Text of embedded speech that is overlapped is in blue and surrounded by orange speaker ID and square brackets.]
LAUGHTER		
<EVENT DESC=LAUGH> or <EVENT DESC=LAUGH WHO=S2>	All laughter is marked. Speaker ID not marked if current speaker laughs.	<LAUGH>, <S8 LAUGH> <SS LAUGH>, etc.

CONTEXTUAL EVENTS		
<EVENT DESC="WRITING ON BOARD"> <EVENT DESC="APPLAUSE"> <EVENT DESC="AUDIO DISTURBANCE">, <EVENT DESC="BACKGROUND NOISE"> <EVENT DESC="SOUND EFFECT">, <EVENT DESC="GASP">	Various contextual (non-speech) events are noted, usually only when they affect comprehension of the surrounding discourse.	<WRITING ON BOARD> <APPLAUSE> <AUDIO DISTURBANCE>, <BACKGROUND NOISE> <SOUND EFFECT>, <GASP>
READING PASSAGES		
<SEG TYPE="READING">.....</SEG>	Used when part of an utterance is read verbatim.	<READING>.....</READING>
FOREIGN WORDS		
<FOREIGN>.....</FOREIGN>	Used for non-English words or phrases.	Italics e.g.: the mother says *c'est quoi*? and Annika says to *parce que* eh and then,...
PRONUNCIATION VARIATIONS		
<SEG TYPE="PRON" SUBTYPE="/seltik/">Celtic</SEG>	Used when an unexpected pronunciation is used that would affect comprehension of the surrounding discourse. Dialect or other phonological variations are generally not represented.	Pronunciation guide follows the word e.g.: ...they asked the librarian for pictures of old Celtic <PRON: /seltik/> uniforms the basketball team, and it turns out that the project was he was supposed to find Celtic <PRON: /keltik/> costumes.
<SIC>...</SIC>	Used when a speaker makes a mistake without self-correcting, and the error might otherwise appear to be a transcribing error.	*(sic)* follows the word. e.g.: despite the fact that that was the era of Women's Liberation like i say on the cover of Newsweek, and Gloria Steinman *(sic)* and uh Betty Friedan...

UNCERTAIN or UNINTELLIGIBLE SPEECH		
(xx) (words) [See 3.3 below for the treatment of this in the POS-tagged version]	Two x's in parentheses indicate one or more words that are completely unintelligible. Words surrounded by parentheses indicate the transcription is uncertain.	i don't (xx) whole (xx) analysis it just struck me... lemme not write it that way (lest it be confused) with C syntax...
NAMES		
When participants' names occur in a recording, they are changed to pseudonyms in the transcript, except in the case of most public colloquia (i.e. COL-prefixed files). In some cases, names of non-present people referred to in the recording are also changed. There is no markup for names.		

3.3 Differences between the versions of MICASE

A version became available via CD-ROM or downloadable zip file in July 2003 for a nominal fee. This newer version incorporates some updates, listed below, which are not reflected in the on-line version:

3.3.2 Differences between the on-line version and the CD-ROM/ distributed version

- the encoding format has been converted from SGML to XML
- some typographical and other errors have been corrected
- spelled words no longer have hyphens between the letters being read out

In addition, please note the following:

- the web version contains extraneous SGML mark-up that speeds up the on-line search engine
- the speech for some speakers has been hidden/deleted, as per consent restrictions, in the web version

3.3.3 Differences between the untagged and tagged versions

In the part-of-speech tagged version of MICASE (not publicly available yet) there are some other differences in mark-up. These changes were all in the direction of XMLizing all mark-up and non-text phenomena in order to ease processing by the CLAWS tagger. The changes are listed below:

- unintelligible speech (in parentheses in the untagged version) is represented by XML in the tagged version:

Untagged MICASE	Tagged MICASE
(xx)	<GAP DESC="UNINTELLIGIBLE"/>
(somethingorother)	<SEG TYPE="UNCLEAR">somethingorother</SEG>

- false starts (indicated by a trailing underscore character in the untagged version) are represented by XML in the tagged version:

Untagged MICASE	Tagged MICASE
`so there will be_ watch this okay`	`so there will be<STRUNC/> watch this okay`

- truncated words (followed by a hyphen in the untagged version) are enclosed within an XML element in the tagged version (NOTE: this means they no longer contribute to word counts and are not part of the flow of the text, unless specifically included):

Untagged MICASE	Tagged MICASE file
`somethingoroth-`	`<GAP DESC="somethingoroth-" REASON="TRUNC"/>`

- pauses are all consistently indicated by XML mark-up in the tagged version, as shown earlier in Table 3-2.

3.3.3.1 Spelling conventions and tokenization in the tagged version

In addition to the above differences in mark-up, some differences in the tagged version of MICASE have to do with the representation and analysis of 'words' in the transcripts:

- **spelled words** and **acronyms/initialisms** (shown with hyphens between letters in the untagged version) appear without hyphens in the tagged version. The former are represented as sequences with spaces between the letters (e.g., `B-A-U-D-R-I-L-L-A-R-D` is `B A U D R I L L A R D` in the tagged version) while the latter appear as 'words' in their own right, unhyphenated (e.g., `E-L-I` is `ELI`; `C-U-N-O-three` is `CUNO-three`).
- **trailing enclitics** and "**fused forms**" have been re-tokenized by the tagger. Most of the words which are typically joined together in speech and spelled as they are pronounced (e.g., *he'll, let's, shouldn'ta*, *wanna*) have been **split up** into their 'component parts' in the tagged version. This affects mainly the modal verbs, but also forms such as *wanna, kinda* and *sorta*. For example, *shouldn'ta* is transcribed as a single orthographic 'word' in the untagged version, but is re-tokenized and tagged as *should*_VM *n't*_XX *'ve*_VHI in the tagged version (i.e., as consisting of three 'morphemes': modal + negative + contracted infinitival *have*).
  A list of the enclitics and fused forms recognized and treated as two or more words by CLAWS can be found at the following URL, where the fused forms used in the British National Corpus (World Edition) are listed exhaustively:
  http://www.comp.lancs.ac.uk/ucrel/bnc2/fused.htm
  This list is not exhaustive for MICASE, however, and many of them do not apply to American speech.

- in a reverse direction to the above tokenisation rules, **multi-word expressions** of various kinds are treated as a **single** grammatical unit by the tagger. For example, *for the most part* is tagged as a four-part multi-word adverb: *for*_RR41 *the*_RR42 *most*_RR43 *part*_RR44 (where 'RR' is the tag for 'general adverb', '41' means 'four parts, first part', '42' means 'four parts, second part', and so on). Similarly, *so that* is tagged as a two-part subordinating conjunction, *so*_CS21 *that*_CS22, and *instead of* is tagged as a two-part preposition, *instead*_II21 *of*_II22. A fuller account of how multi-words are treated by CLAWS may be found in the manual which accompanies the tagged version of the British National Corpus (World Edition) at http://www.comp.lancs.ac.uk/ucrel/bnc2/bnc2postag_manual.htm (see the section on "Tokenization: splitting the text into words"). A full list of multi-word expressions so treated by the tagger is also available from that site.

Users of the tagged version of the corpus should familiarize themselves with how 'words' are tokenized, as it affects how search terms are specified when concordancing.

Chapter 4

Statistical overviews and word counts

The four tables provided here give descriptive frequency figures for MICASE, showing the composition of the corpus by speaker and speech event categories. These word counts are for the untagged version of MICASE, and 'words' were defined as strings bounded by spaces or punctuation; hyphenated words (e.g., *eighty-five, cross-sectional, non-native*) and words with apostrophes (e.g., *she'll, shouldn't've, don't, John's*) counted as one word.

Table 4-1 Speaker and word counts by speaker categories

Speaker Category			Total Speakers	Total Words	% of Total Corpus
Gender	Male		**729**	**786,487**	**46%**
	Female		**842**	**909,053**	**54%**
Academic Role*	Faculty		**160**	**825,829**	**49%**
		[Male	*84*	*446,925*	*26%]*
		[Female	*76*	*378,904*	*22%]*
	Students		**1,039**	**742,348**	**44%**
		Undergraduates	782	368,433	22%
		[Male	*336*	*142,102*	*8%]*
		[Female	*446*	*226,331*	*13%]*
		Graduates	257	373,915	22%
		[Male	*121*	*158,696*	*9%]*
		[Female	*136*	*215,219*	*13%]*

Language Status	Native Speakers	**1,449**	**1,493,586**	**88%**
	Non-native speakers	**122**	**201,954**	**12%**
Totals		1,571	1,695,540	

Table 4-2 Speaker and word counts by academic division

Academic Division	Speech Events	Speakers	Words	% of Total Corpus	% Male	% Female	% Faculty*	% Students*
Humanities & Arts	36	349	434,669	26	56	44	63	29
Social Sciences & Education	35	452	420,347	25	37	63	44	55
Biological & Health Sciences	32	257	325,456	19	41	59	55	42
Physical Sciences & Engineering	36	314	358,776	21	55	45	44	52
Other/NA	13	199	156,292	9	37	63	20	41
Totals	152	1,571	1,695,540					

Table 4-3 Speaker and word counts by primary discourse mode

Primary Discourse Mode	Speech Events	Speakers	Words	% of Total Corpus	% Male	% Female	% Faculty *	% Students*
Monologic	61	472	554,335	33	50	50	84	14
Panel	9	133	141,505	8	27	73	16	76
Interactive	57	643	715,333	42	46	54	26	63
Mixed	25	323	284,367	17	51	49	54	39
Totals	152	1,571	1,695,540					

Table 4-4 Speaker and word counts by speech event type

Speech Event Type	Tran-scripts	Speak-ers	Words	% of Total Corpus	% Male	% Female	% Fac-ulty*	% Stu-dents*
Advising	5	20	58,817	3.5	43.5	56.5	14.2	37.2
Colloquia	13	118	151,639	8.9	52.9	47.1	76.9	10.8
Discussion Sections	9	112	74,904	4.4	36.8	63.2	33	66.7
Diss. Defenses	4	26	56,837	3.4	55.1	44.9	36.5	62.7
Interviews	3	6	13,015	0.8	82.6	17.4	56.0	44.0
Labs	8	42	73,815	4.4	69.8	30.2	15.1	67.9
Large Lectures	31	217	257,311	15.2	52.6	47.4	93.5	5.9
Small Lectures	31	289	320,893	18.9	43.8	56.2	74.0	22.6
Meetings	6	60	70,038	4.1	65.8	34.2	15.8	61.6
Office Hours	8	79	120,629	7.1	32.1	67.9	26.9	72.8
Seminars	8	79	151,071	8.9	60.2	39.8	58.8	34.9
Study Groups	8	36	129,725	7.7	31.7	68.3	0	100.0
Student Presentations	11	146	143,369	8.5	23.9	76.1	15.4	77.6
Service En-counters	2	90	24,691	1.5	40.6	59.4	.02	60.2
Tours	2	19	21,768	1.3	58.4	41.6	0	60.9
Tutorials	3	18	27,014	1.6	35.4	64.7	15.9	80.9

*** Note**: In these tables, percentages for faculty and students do not add up to 100% because of other speaker roles (e.g. staff, researchers, visitors) not included in these counts.

Chapter 5

Using MICASE On-line

MICASE was designed to be freely available to as many researchers, instructors and students as possible. To that end, since May, 2002 the entire corpus has been accessible on the web, on a site with a searchable interface much like a concordance program. The MICASE On-line search engine homepage is at: http://www.hti.umich.edu/m/micase/ This chapter provides help with using this on-line interface.

There are two modes for exploring MICASE On-line: **Browse** and **Search**. The two modes are described very briefly here, followed by more specific directions and suggestions for browsing and searching the corpus.

5.1 Browse MICASE

The browse function returns lists of *transcripts* that match your criteria. *(Please see Sections 2.3 and 2.4 for explanations of speech event and speaker attributes.)*

You can specify

- speech event parameters (type of event, academic division, academic discipline, participant level, and discourse mode);
- characteristics (gender, age, academic position/role, and native language) of any speaker *participating* in the event (not the speaker uttering the word or phrase of interest, if specified. See boxed note under "Search MICASE" below); and/or
- a word or phrase of interest.

The list of transcripts returned by the browse function includes each file's ID, transcript name, the number of occurrences of the key word/phrase (if specified), recording length, and word count.

From this list, you can

- access the transcript of interest by clicking on the file ID in the left-hand column, and then clicking again on "View entire transcript in HTML."

5.2 Search MICASE

The search function returns a list of each *occurrence* of the specified word or phrase in concordance format (Key Word in Context). It is possible to search the entire corpus for such occurrences, or you can restrict the search using any speech event or speaker category. *(Please see Sections 2.3 and 2.4 for explanations of speech event and speaker attributes.)*

You must first specify

- the word or phrase of interest.

You can also specify

- attributes of the speech event (type of event, academic division, academic discipline, participant level, and discourse mode) in which the utterance occurs,
- characteristics of the speaker (gender, age group, academic position/role, and native language) uttering the key word or phrase (See boxed note below), and/or
- a word or phrase that must appear within a specified proximity to the search word(s).

You can also select

- up to two speaker attributes (gender, age group, academic role, and native speaker status) that you wish to *appear* in the results display. (If selected, you can subsequently sort results based on these categories.) **Note**: These do not have to be the same categories used to narrow your search. In fact, it can be redundant to select a speaker attribute to appear in the results column if you have already specified that attribute as a search criterion. For example, if you restrict your search to include only female speakers, and then check to display the "gender" attribute in the results, it will display the gender of each speaker, which we already know is female.

> Note the difference between selecting speakers in browse vs. search modes: In browse mode, the selected speaker(s) includes *any* speaker participating in the event. In search mode, the selected speaker is the person uttering the search word or phrase.

The "hit list"/concordance results screen displays the total number of matches, the file ID, the speaker ID, and speaker attribute(s), if selected for display.

From this list, you can

- access the transcript of interest by clicking on the file ID in the left-hand column, and then clicking again on "View Entire Transcript in HTML,"
- display the full utterance by clicking on the concordance line number (from which point you can then access the transcript in which the utterance appears),
- access information about the speaker (gender, age, and academic position/role) by clicking on the speaker ID, and
- sort the results according to the words appearing to the left or right of the key word(s), and/or according to the speaker attributes selected for display (see above).

5.3 Tips

IN EITHER MODE	
To select more than one item in a speech event or speaker attribute category	Hold down control key while clicking on additional selections.
IN BROWSE MODE	
To see a list of all available transcripts	Leave speech event and speaker attribute categories set to the default "All."
To see a list of only the transcripts that match certain criteria	Select the type of speech event and speaker attribute(s) of interest.
To see a list of transcripts containing a particular word or phrase, and the number of occurrences in each transcript	Specify the word or phrase in the optional search section at the bottom of the browse page, and leave all categories set to "All."
To see a list of transcripts that match certain criteria AND contain a particular word or phrase (includes the number of occurrences in each transcript)	Select the type of speech event and speaker attribute(s) of interest, and specify the word or phrase in the optional search section at the bottom of the browse page.

IN SEARCH MODE	
Types of searches allowed	Single words or multi-word phrases Single word or phrase plus a context word, i.e., another word or phrase found within a specified proximity of the main search term The wildcard character * may be used at the end (but not the beginning) of a search word to represent zero or more characters (e.g., typing in *walk** will give you *walk, walks, walked, walker, walkers,* and *walking*). These characters appear separately to the right of the search word and can be sorted. Current limitations: The search engine does not currently accept Boolean searches with "OR" (i.e., you cannot search for all instances of two or more terms, e.g., you cannot search for "*woman* OR *women*").
To see concordance lines for any word or phrase used by *any* speaker in *all* transcripts	Enter a word or phrase in the first box of the search page, and leave all speaker and speech event attributes set to "All."
To see concordance lines for any word or phrase used by *any* speaker in *specific types* of speech events only	Enter a word or phrase in the first box of the search page, and select the desired speech event attributes in the first column. Select up to 2 speaker attributes that you wish to *appear* in the results display. If selected, results can be sorted accordingly.
To see concordance lines for any word or phrase used by *specific* speakers in *all* transcripts	Enter a word or phrase in the first box of the search page, and select the desired speaker attributes in the far right column. Select up to 2 speaker attributes that you wish to *appear* in the results display. **Hint:** It is redundant to select a speaker attribute to appear in the results column if you have already specified that attribute as your search criterion. See 5.2 above.
To see concordance lines for any word or phrase used by *specific* speakers in *specific types* of speech events only	Enter a word or phrase in the first box of the search page. Select the desired speech event attributes in the first column to the right. Select the desired speaker attributes in the far right column. Select up to 2 speaker attributes that you wish to *appear* in the results display. **Hint:** It is redundant to select a speaker attribute to appear in the results column if you have already specified that attribute as your search criteria. See 5.2 above.

To re-sort results	Use pull-down menus at top of search results page to specify primary (1), secondary (2), or tertiary (3) sort filters. Sort options include: 1st-5th word to the left or right (1L-5L, 1R-5R) The search term itself (e.g. in cases where the word may or may not be followed by a punctuation mark) The speaker attribute(s) chosen to appear in the search results. See 5.2 above. **NOTE: Results returning more than 500 hits cannot be sorted**.
To view entire utterance in which search term appears	Click on the concordance line number on the left of the Key Word in Context line.
To see speaker attributes for an utterance	Click on speaker ID in the right column.
IN EITHER MODE	
To view entire transcript	Click on File ID in the far left column; this takes you to the header view. At the bottom of the header view page, click on the "View entire transcript in HTML" link.
HTML Transcription Conventions	Search term is highlighted **in red**. Context term (if specified in Search Mode) is highlighted **in green**. **Blue text** indicates overlapping speech. **Orange text** indicates embedded turn (different speaker). <L> = laugh <P> = pause longer than 3 seconds <E> = contextual or non-verbal event (identified in utterance and full transcript views.) For a full description of all transcription conventions, please see section 3.2
Current limitations	Note that limiting your search by too many speech event or speaker parameters may result in a very small number of hits, or none at all. Results returning more than 500 hits cannot be sorted. If you discover something that doesn't seem to be working as it should, please let us know by sending a message to micase@umich.edu. We would also appreciate suggestions or comments as to ways we could improve the site or the functionality of the on-line search features.

Chapter 6

Distribution, availability, and copyright restrictions

There are two different ways to do research on MICASE. If you have obtained the SGML or XML files, you can use them with your own software on your own computer. Otherwise, you are welcome to use MICASE On-line, our web-based search facility, linked from the MICASE home page. In either case, please familiarise yourself with our copyright restrictions.

6.1 Obtaining a copy of the corpus

From the MICASE On-line site, it is possible to download individual MICASE transcripts and thus obtain a copy of the whole corpus. This involves going through 2 or 3 steps/clicks per transcript to obtain all 152 files. If you want to obtain MICASE by this method, go to the Transcript Header View of the first file you are interested in (by clicking on the filename on the "results" page of either a "Browse" or a "Search"), then choose "Download entire transcript in SGML." Repeat this procedure for all the other transcripts. Do note, however, the differences between this on-line version and the CD-ROM distributed version, as listed in section 3.3.2.

If you want to do independent research on MICASE using your own tools or concordance package (e.g., Wordsmith or MonoConc Pro) and want the full, complete texts, you may purchase a single-user or site license for the corpus for a nominal fee. This can be done through our web site using our secure on-line order form or by filling out the form in Appendix D.

6.2 Availability of sound files

In the future, we will be distributing some of the original sound files, compressed in MP3 format. These sound files will be available to bona fide academic researchers who agree to the terms of our license, especially regarding anonymity issues and use of the materials for commercial purposes. We also have plans to make some of the sound files available on the web in RealAudio format. For both formats, however, some speaker consent restrictions will prevent us from making the complete set of recordings publicly available. We will post updates on the availability of these sound files on our web site.

6.3 Copyright restrictions and licenses

Obtaining the MICASE corpus implies that you accept the following copyright and license terms:

MICASE copyright statement

The MICASE project is owned by the Regents of the University of Michigan, who hold the copyright. The database has been developed by the English Language Institute, and the web interface by Digital Library Production Services. The original DAT audiotapes are held in the English Language Institute and may be consulted by bona fide researchers under special arrangements. The database is freely available at the MICASE website for study, teaching, and research purposes, and copies of the transcripts may be distributed, as long as either this statement of availability or the citation given below appears in the text. However, if any portion of this material is to be used for commercial purposes, such as for textbooks or tests, permission must be obtained in advance and a license fee may be required. Furthermore, some restrictions apply on the citation of specific portions of some of the transcripts in educational presentations and publications; all such restrictions are noted in the headers of individual files of the corpus. For further information about copyright permission, please contact Dr. Sarah Briggs at slbriggs@umich.edu. The recommended citation for MICASE is: R. C. Simpson, S. L. Briggs, J. Ovens, and J. M. Swales. (1999) The Michigan Corpus of Academic Spoken English. Ann Arbor, MI: The Regents of the University of Michigan.

MICASE individual and site licenses

An *Individual License* grants one named person permission to use the MICASE corpus for non-commercial purposes. The corpus may not be redistributed to non-licensed users and may not be used in commercial applications, materials or publications, as stated in the MICASE copyright statement, without the express permission of the copyright holder.

A *Site License* grants up to 25 users at the licensed institution permission to use the MICASE corpus for non-commercial purposes. The corpus may not be redistributed to non-licensed users and may not be used in commercial applications, materials or publications, as stated in the MICASE copyright statement, without the express permission of the copyright holder. All users of the corpus at the licensed institution must be notified of the terms of this license.

.

Appendix A
Index of Transcripts

File ID	File name	# of speakers	Recording length (mins.)	Transcript word count
adv105su068	American Culture Advising	5	42	8135
adv285sg135	Graduate Education Advising	2	54	8829
adv355su094	Linguistics Independent Study Advising	2	52	6579
adv700ju023	Honors Advising	4	52	9209
adv700ju047	Academic Advising	7	124	26065
col140mx114	Peking Opera Colloquium	13	74	11299
col200mx133	Chemical Biology Colloquium	10	63	9941
col285mx038	Education Colloquium	9	52	7898
col385mu054	Public Math Colloquium	3	51	7236
col425mx075	Ecological Agriculture Colloquium	8	97	17061
col475mx082	Philosophy Colloquium	13	95	15232
col485mx069	Nobel Laureate Physics Lecture	9	87	14476
col605mx039	Women's Studies Guest Lecture	5	65	9386
col605mx132	Christianity and the Modern Family Colloquium	14	78	11027
col999mg053	Career Planning and Placement Workshop	12	76	11324
col999mx036	Provost Public Lecture	4	61	8823
col999mx040	Women in Science Conference Panel	11	105	18671
col999mx059	Problem Solving Colloquium	7	66	9265
def270sf061	Artificial Intelligence Dissertation Defense	6	113	20621
def305mx131	Fossil Plants Defense	9	57	9558
def420mx022	Music Dissertation Defense	6	91	14982
def500sf016	Social Psychology Dissertation Defense	5	76	11676
dis115ju087	Intro Anthropology Discussion Section	17	51	7893
dis150ju130	Intro Astronomy Discussion Section	4	33	5338
dis175ju081	Intro Biology Discussion Section	22	59	6899
dis175su027	Biology of Birds Discussion Section	11	55	7424
dis195su117	Heat and Mass Transfer Discussion Section	4	48	7570
dis280su058	Economics Discussion Section	8	61	8526

File ID	File name	# of speakers	Recording length (mins.)	Transcript word count
dis315ju101	History Review Discussion Section	19	119	15679
dis475mu012	Philosophy Discussion Section	9	51	8355
dis495ju119	Intro to American Politics Discussion Section	18	55	7220
int175sf003	Interview with Botanist	2	31	5016
int425jg001	Graduate Student Research Interview 1	2	34	5063
int425jg002	Graduate Student Research Interview 2	2	20	2936
lab175su026	Biology of Birds Field Lab	0	92	10696
lab175su032	Biology of Fishes Field Lab	3	89	10617
lab175su033	Biology of Fishes Lab	1	95	7246
lab200ju018	Chemistry Lab	2	47	7511
lab205su045	Hydraulics Problem Solving Lab	10	78	9457
lab500su044	Biopsychology Lab	9	52	8847
lab500su089	Cognitive Psychology Research Lab	12	82	13623
lab575ju095	Intro Statistics Lab	5	47	5818
lel105su113	History of the American Family Lecture	9	81	10621
lel115ju090	Intro Anthropology Lecture	2	74	11206
lel115su005	Medical Anthropology Lecture	5	69	11470
lel115su107	Race and Human Evolution Lecture	8	77	9395
lel140su074	Japanese Literature Lecture	7	44	8213
lel175ju086	Practical Botany Lecture	1	48	3788
lel175ju112	General Ecology Lecture	6	51	4331
lel175ju154	Intro to Evolution Lecture	3	98	11699
lel175mu014	Intro Biology First Day Lecture	7	47	6613
lel175su098	Intro to Biochemistry Lecture	11	82	7917
lel175su106	Biology of Cancer Lecture	2	70	10964
lel185su066	Behavior Theory Management Lecture	49	80	12698
lel195su120	Separation Processes	8	48	4277
lel200ju105	Inorganic Chemistry Lecture	1	50	6532
lel200mu110	Structure and Reactivity II Lecture	8	45	3815
lel215su150	Sports and Daily Life in Ancient Rome Lecture	1	71	12581
lel220ju071	Intro Communication Lecture	2	76	7682
lel220su073	Media Impact in Communication Lecture	13	72	9164
lel280jg051	Graduate Macroeconomics Lecture	7	76	7865
lel295ju035	Intro Engineering Lecture	6	52	5563
lel300su020	Literature and Social Change Lecture	3	84	9568
lel300su076	Fantasy in Literature Lecture	7	83	12682

File ID	File name	# of speakers	Recording length (mins.)	Transcript word count
lel305ju092	Intro Oceanography Lecture	10	63	8175
lel320ju143	Renaissance to Modern Art History Lecture	2	50	7951
lel320ju147	Twentieth Century Arts	4	41	5981
lel485ju097	Intro to Physics Lecture	1	49	7220
lel500ju034	Intro Psychology Lecture	1	47	7266
lel500su088	Drugs of Abuse Lecture	5	68	9623
lel542su096	Perspectives on the Holocaust Lecture	9	100	6521
lel565su064	Principles in Sociology Lecture	16	82	10246
lel575mx055	Golden Apple Award Statistics Lecture	3	45	5688
les115mu151	Archeology of Modern American Life Lecture	16	73	10463
les165jg121	Rehabilitation Engineering and Technology	9	49	7025
les175su025	Biology and Ecology of Fishes Lecture	6	70	9326
les175su028	Biology of Birds Lecture	7	84	11417
les175su031	Biology of Fishes Group Activity	1	19	2705
les175su079	Microbial Genetics Lecture	8	85	12264
les205jg124	Intro to Groundwater Hydrology Lecture	11	82	13334
les215mu056	Intro Latin Lecture	9	50	5460
les220su140	Ethics Issues in Journalism Lecture	24	83	15305
les235su099	Intro Programming Lecture	6	50	7623
les280jg138	Labor Economics Lecture	6	77	11822
les300su103	American Literature Lecture	14	99	15385
les305mu108	Dynamic Earth Lecture	9	51	6753
les315su129	African History Lecture	10	68	8755
les320su085	Visual Sources Lecture	12	69	11493
les330jg052	Graduate Industrial Operations Engineering Lecture	6	81	10096
les335jg065	Graduate Online Search and Database Lecture	24	147	18039
les355su009	Historical Linguistics Lecture	6	69	12569
les385su007	Number Theory Math Lecture	10	36	3585
les405jg078	Graduate Cellular Biotechnology Lecture	4	83	12584
les420mg134	Beethoven Lecture	11	75	7161
les425jg077	Graduate Population Ecology Lecture	4	44	5041
les425su093	Spring Ecosystems Lecture	7	74	10678
les445su067	Radiological Health Engineering Lecture	10	98	12876
les485mg006	Graduate Physics Lecture	7	105	13008
les495ju063	Political Science Lecture	6	96	14806

File ID	File name	# of speakers	Recording length (mins.)	Transcript word count
les500ju136	Honors Intro Psychology	4	49	5531
les500su102	Intro to Psychopathology Lecture	2	52	7938
les565mx152	Statistics in Social Sciences Lecture	10	109	15432
les565su137	Sex, Gender, and the Body Lecture	20	73	13153
les605su080	Women in the Bible Lecture	10	75	9266
mtg270sg049	Artificial Intelligence Research Group Meeting	7	94	16760
mtg400mx008	Immunology Lab Meeting	6	60	8394
mtg425jg004	Natural Resources Research Group Meeting	5	83	9130
mtg485sg142	Physics Research Group Meeting	8	41	8504
mtg999st015	Forum for International Educators Meeting	11	102	14910
mtg999su043	Student Government Meeting	23	66	12340
ofc115su060	Anthropology of American Cities Office Hours	6	178	29635
ofc175ju145	Intro Biology Exam Review	9	55	8221
ofc195su116	Heat and Mass Transfer Office Hours	23	137	19498
ofc270mg048	Computer Science Office Hours	11	116	19044
ofc280su109	Economics Office Hours	10	92	13114
ofc300ju149	Intro to Poetry Office Hours	4	88	11638
ofc320su153	Art History Office Hours	5	66	8971
ofc575mu046	Statistics Office Hours	11	52	10508
sem140jg070	Graduate Buddhist Studies Seminar	8	167	21537
sem300mu100	Introduction to Composition	11	125	20113
sem340jg072	Graduate Public Policy Seminar	11	143	24180
sem365vo029	Professional Mechanical Engineering Seminar	7	90	12445
sem475ju084	First Year Philosophy Seminar	13	72	10300
sem475mx041	Graduate Philosophy Seminar	3	125	20979
sem495su111	Politics of Higher Education	19	108	18687
sem545mg083	Graduate French Cinema Seminar	7	169	22830
sgr175mu126	Intro Biology Study Group	5	103	22422
sgr175su123	Biochemistry Study Group	4	109	16192
sgr195su127	Chemical Engineering Group Project Meeting	4	77	10386
sgr200ju125	Organic Chemistry Study Group	6	101	16377
sgr385su057	Math Study Group	3	132	16130
sgr565su144	American Family Group Project Meeting	5	85	13388
sgr999mx115	Objectivism Student Group	5	125	20830

File ID	File name	# of speakers	Recording length (mins.)	Transcript word count
sgr999su146	Undergrad. Social Science Study Group	4	64	14000
stp095su139	Black Media Student Presentations	13	66	9282
stp125jg050	Architecture Critiques	12	123	22596
stp165jg122	Rehab Engineering and Technology Student Presentations	5	32	5241
stp175su141	Teaching Biochemistry Student Presentations	18	121	17224
stp200ju019	Chemistry Discussion Section Student Presentations	18	51	6575
stp285su013	Multicultural Issues in Education Student Presentations	11	72	12354
stp355mg011	Bilingualism Student Presentations	7	99	15153
stp355su010	Second Language Acquisition Student Presentations	12	69	9791
stp450sg128	Nursing Student Presentations	10	155	22223
stp545ju091	Brazilian Studies Student Presentations	14	78	12345
stp560jg118	Community Change Student Presentations	26	66	10585
svc999mx104	Media Union Service Encounters	79	187	17093
svc999mx148	Science Learning Center Service Encounters	11	121	7598
tou999ju030	Freshman Orientation Tour	11	77	12995
tou999mx062	Art Museum Tour	8	66	8773
tut150mu042	Astronomy Peer Tutorial	12	102	19763
tut301mu021	English Composition Tutorial	4	45	3303
tut578mx037	Technical Communications Tutorial	2	25	3948

Appendix B
The CLAWS C8++ Tag set

[* Tags which are different from the C7 tag set are in bold]

APPGE	possessive pronoun, pre-nominal (e.g. "my", "your", "our")
AT	article (e.g. "the", "no")
AT1	singular article (e.g. "a", "an", "every")
BCL	before-clause marker (e.g. "in order (that)", "in order (to)")
CC	coordinating conjunction (e.g. "and", "or")
CCB	adversative coordinating conjunction ("but")
CS	subordinating conjunction (e.g. "if", "because", "unless", "so", "for")
CSA	"as" (as conjunction)
CSN	"than" (as conjunction)
CST	"that" (as conjunction)
CSW	"whether" (as conjunction)
DA	after-determiner or post-determiner capable of pronominal function (e.g. "such", "former", "same")
DA1	singular after-determiner (e.g. "little", "much")
DA2	plural after-determiner (e.g. "few", "several", "many")
DAR	comparative after-determiner (e.g. "more", "less", "fewer")
DAT	superlative after-determiner (e.g. "most", "least", "fewest")
DB	before determiner or pre-determiner capable of pronominal function ("all", "half")
DB2	plural before-determiner ("both")
DD	determiner (capable of pronominal function) (e.g. "any", "some")
DD1	singular determiner (e.g. "this", "that", "another")
DD2	plural determiner ("these", "those")
DDL	**wh-determiner, relative ("which")**
DDLGE	**wh-determiner, relative, genitive ("whose")**
DDQ	wh-determiner, interrogative ("which", "what")
DDQGE	wh-determiner, interrogative, genitive ("whose")
DDQV	wh-ever determiner, interrogative ("whichever", "whatever")
EX	existential "there"
FO	formula
FU	unclassified word
FW	foreign word
GE	germanic genitive marker - (' or 's)
IF	"for" (as preposition)
II	general preposition

IO	"of" (as preposition)
IW	"with", "without" (as prepositions)
JJ	general adjective
JJR	general comparative adjective (e.g. "older", "better", "stronger")
JJT	general superlative adjective (e.g. "oldest", "best", "strongest")
JK	catenative adjective ("able" in "be able to", "willing" in "be willing to")
MC	cardinal number, neutral for number ("two", "three"..)
MC1	singular cardinal number ("one")
MC2	plural cardinal number (e.g. "sixes", "sevens")
MCGE	genitive cardinal number, neutral for number ("two's", "100's")
MCMC	hyphenated number ("40-50", "1770-1827")
MD	ordinal number (e.g. "first", "second", "next", "last")
MF	fraction, neutral for number (e.g. "quarters", "two-thirds")
ND1	singular noun of direction (e.g. "north", "southeast")
NN	common noun, neutral for number (e.g. "sheep", "cod", "headquarters")
NN1	singular common noun (e.g. "book", "girl")
NN2	plural common noun (e.g. "books", "girls")
NNA	following noun of title (e.g. "M.A.")
NNB	preceding noun of title (e.g. "Mr.", "Prof.")
NNL1	singular locative noun, in naming expression (e.g. "Island", "Street")
NNL2	plural locative noun (e.g. "Islands", as in "Virgin Islands")
NNO	numeral noun, neutral for number (e.g. "dozen", "hundred")
NNO2	numeral noun, plural (e.g. "hundreds", "thousands")
NNT1	temporal noun, singular (e.g. "day", "week", "year")
NNT2	temporal noun, plural (e.g. "days", "weeks", "years")
NNU	unit of measurement, neutral for number (e.g. "in", "cc")
NNU1	singular unit of measurement (e.g. "inch", "centimetre")
NNU2	plural unit of measurement (e.g. "ins.", "feet")
NP	proper noun, neutral for number (e.g. "IBM", "Andes")
NP1	singular proper noun (e.g. "London", "Jane", "Frederick")
NP2	plural proper noun (e.g. "Browns", "Reagans", "Koreas")
NPD1	singular weekday noun (e.g. "Sunday")
NPD2	plural weekday noun (e.g. "Sundays")
NPM1	singular month noun (e.g. "October")
NPM2	plural month noun (e.g. "Octobers")
PN	indefinite pronoun, neutral for number ("none")
PN1	indefinite pronoun, singular (e.g. "anyone", "everything", "nobody", "one")
PNLO	**objective wh-pronoun, relative ("whom")**
PNLS	**subjective wh-pronoun, relative ("whom")**
PNQO	objective wh-pronoun, interrogative ("whom")
PNQS	subjective wh-pronoun, interrogative ("who")
PNQV	wh-ever pronoun ("whoever")
PNX1	reflexive indefinite pronoun ("oneself")
PPGE	nominal possessive personal pronoun (e.g. "mine", "yours")

PPH1	3rd person sing. neuter personal pronoun ("it")
PPHO1	3rd person sing. objective personal pronoun ("him", "her")
PPHO2	3rd person plural objective personal pronoun ("them")
PPHS1	3rd person sing. subjective personal pronoun ("he", "she")
PPHS2	3rd person plural subjective personal pronoun ("they")
PPIO1	1st person sing. objective personal pronoun ("me")
PPIO2	1st person plural objective personal pronoun ("us")
PPIS1	1st person sing. subjective personal pronoun ("I")
PPIS2	1st person plural subjective personal pronoun ("we")
PPX1	singular reflexive personal pronoun (e.g. "yourself", "itself")
PPX2	plural reflexive personal pronoun (e.g. "yourselves", "themselves")
PPY	2nd person personal pronoun ("you")
RA	adverb, after nominal head (e.g. "else", "galore")
REX	adverb introducing appositional constructions ("namely", "e.g.")
RG	degree adverb ("very", "so", "too")
RGQ	wh- degree adverb ("how")
RGQV	wh-ever degree adverb ("however")
RGR	comparative degree adverb ("more", "less")
RGT	superlative degree adverb ("most", "least")
RL	locative adverb (e.g. "alongside", "forward")
RP	prep. adverb, particle (e.g. "about", "in")
RPK	prep. adv., catenative ("about" in "be about to")
RR	general adverb
RRQ	wh- general adverb ("where", "when", "why", "how")
RRQV	wh-ever general adverb ("wherever", "whenever")
RRR	comparative general adverb (e.g. "better", "longer")
RRT	superlative general adverb (e.g. "best", "longest")
RT	quasi-nominal adverb of time (e.g. "now", "tomorrow")
TO	infinitive marker ("to")
UH	interjection (e.g. "oh", "yes", "um")
	[The following three tags are unique to MICASE:]
UHY	**interjection, positive (e.g. *uh-huh, yuh-huh*)**
UHN	**interjection, negative (e.g. *uh-uh, huh-uh, m-m, hm-m, nuh-uh*)**
UHE	**interjection, exclamatory (e.g. *uh-oh, oh-oh, oops*)**
VAB0	**base form of verb "BE" (auxiliary), imperative or subjunctive**
VABDR	**"were" (auxiliary)**
VABDZ	**"was" (auxiliary)**
VABG	**"being" (auxiliary)**
VABI	**"be" infinitive (auxiliary)**
VABM	**"am" (auxiliary)**
VABN	**"been" (auxiliary)**
VABR	**"are" (auxiliary)**
VABZ	**"is" (auxiliary)**
VVB0	**base form of "BE" (lexical vb), imperative or subjunctive**
VVBDR	**"were" (lexical vb)**

VVBDZ **"was" (lexical vb)**
VVBG **"being" (lexical vb)**
VVBI **"be" infinitive (lexical vb)**
VVBM **"am" (lexical vb)**
VVBN **"been" (lexical vb)**
VVBR **"are" (lexical vb)**
VVBZ **"is" (lexical vb)**
VAD0 **base form of verb "DO" (auxiliary), indicative, imperative or subjunctive**
VADD **"did" (auxiliary)**
VADZ **"does" (auxiliary)**
VVD0 **base form of verb "do" (finite)**
VVDD **"did"**
VVDG **"doing"**
VVDI **"do" infinitive ("I may do..." "To do...")**
VVDN **"done"**
VVDZ **"does"**
VAH0 **base form of "HAVE" (auxiliary), indicative, imperative or subjunctive**
VAHD **"had" (past tense)**
VAHG **"having"**
VAHI **"have" infinitive**
VAHZ **"has"**
VVH0 **base form of verb "HAVE" (finite)**
VVHD **"had" (past tense)**
VVHG **"having"**
VVHI **"have" infinitive**
VVHN **"had" (past participle)**
VVHZ **"has"**
VM modal auxiliary ("can", "will", "would", etc.)
VMK modal catenative ("ought", "used")
VV0 base form of lexical verb (e.g. "give", "work")
VVD past tense of lexical verb (e.g. "gave", "worked")
VVG -ing participle of lexical verb (e.g. "giving", "working")
VVGK -ing participle catenative ("going" in "be going to")
VVI infinitive (e.g. "to give..." "It will work...")
VVN past participle of lexical verb (e.g. "given", "worked")
VVNK past participle catenative (e.g. "bound" in "be bound to")
VVZ -s form of lexical verb (e.g. "gives", "works")
WPR **relative pronoun, "that"**
XX "not", "n't"

[Note: In the current tagged version of MICASE, the question mark is the only punctuation mark used.]

YBL **punctuation tag - left bracket**
YBR **punctuation tag - right bracket**

YCOL	**punctuation tag - colon**
YCOM	**punctuation tag - comma**
YDSH	**punctuation tag - dash**
YEX	**punctuation tag - exclamation mark**
YLIP	**punctuation tag - ellipsis**
YQUE	**punctuation tag - question mark**
YQUO	**punctuation tag - quotes**
YSCOL	**punctuation tag - semicolon**
YSTP	**punctuation tag - full-stop**
ZZ1	singular letter of the alphabet (e.g. "A", "b")
ZZ2	plural letter of the alphabet (e.g. "A's", "b's")

Appendix C
MICASE-based Publications and Presentations

This appendix lists publications and conference presentations which have drawn on the MICASE corpus. In addition to these references, on our website (http://www.lsa.umich.edu/eli/micase/micase.htm) we have several sets of sample pedagogical materials and "Kibbitzers" (snippets of research/language notes for ESL purposes) based on MICASE, under the section "Pedagogical and other forays". A more general bibliography of research related to academic speaking is also posted there.

Publications

Forthcoming

Burke, A. & Swales, J. M. "It's really fascinating work": Differences in evaluative adjectives across academic registers. In Meyer, C. & P. Leistyna (eds.) *Corpus Analysis: Language structure and use.* Amsterdam: Rodopi.

Mauranen, A. "They're a little bit different": Variation in hedging in academic speech. In Aijmer, K. & A-B Stenström (eds.) *Discourse Patterns in Spoken and Written Corpora.* Amsterdam: John Benjamins.

Simpson, Rita C. and Mendis, Dushyanthi. A corpus-based study of idioms in academic speech. *TESOL Quarterly.*

Swales, John M. (to appear in 2004) *Research Genres: Explorations and Applications.* New York: Cambridge University Press. [Chapters 5 and 6 are especially related to MICASE]

2002

Poos, Deanna and Simpson, Rita C. Cross-disciplinary comparisons of hedging: Some findings from the Michigan Corpus of Academic Spoken English. In Reppen, R., Fitzmaurice, S. M., and Biber, D., *Using Corpora to Explore Linguistic Variation.* Amsterdam: John Benjamins. pp.3-23.

Swales, John M. Integrated and fragmented worlds: EAP materials and corpus linguistics. In J. Flowerdew (Ed.) *Academic Discourse* (pp.153-167). London: Longman

Mauranen, Anna. "A good question": Expressing evaluation in academic speech. In Cortese, G. & P. Riley (eds.) *Domain-specific English: Textual practices across communities and classrooms.* Frankfurt: Peter Lang. pp.115-140.

2001

Lindemann, Stephanie & Mauranen, Anna. It's just real messy: The occurrence and function of 'just' in a corpus of academic speech. *English for Specific Purposes*, 20:459-475.

Mauranen, Anna. Reflexive academic talk: Observations from MICASE. In Simpson, R. C. & J. M. Swales (eds.), pp.165-178.

Mauranen, Anna. Descriptions or explanations? Some methodological issues in Contrastive Rhetoric. In M. Hewings (ed.) *Academic Writing in Context: Implications and applications*. Birmingham: The University of Birmingham Press. pp.43-54.

Powell, Christina & Simpson, Rita C. Collaboration between corpus linguists and digital librarians for the MICASE web search interface. In Simpson, R. C. & J. M. Swales (eds.), pp.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 C. & Swales, John M. North American perspectives on corpus linguistics at the millennium. In Simpson, R. C. & J. M. Swales (eds.), 1-14.

Swales, John M. Metatalk in American academic talk: The cases of "point" and "thing." *Journal of English Linguistics*, 29:34-54.

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

2000

Simpson, Rita C., Lucka, Bret & Ovens, Janine. Methodological challenges of planning a spoken corpus with pedagogical outcomes. In Burnard, Lou & Tony McEnery (eds.), *Rethinking Language Pedagogy from a Corpus Perspective*: Papers from the third international conference on Teaching and Language Corpora (TALC), 43-49. Frankfurt: Peter Lang.

Pedagogical materials incorporating MICASE data

Reinhart, Susan M. 2002. *Giving Academic Presentations*. Ann Arbor: University of Michigan Press.

Conference presentations

2003

Lee, David Y.W. Spoken lexicogrammar and discourse patterns in the academy: MICASE past, present and future. Paper presented at Corpus Linguistics 2003, Mar. 2003, Lancaster University, UK.

Mauranen, Anna. "It seems to me like you're saying": Formulae in argumentative discussion. AAAL, Mar. 2003., Arlington, VA.

Nesi, Hilary. Enumeration as a predictive category in academic monologue. AAAL, Mar. 2003, Arlington, VA.

Simpson, Rita C. Academic Spoken English: What Are the Questions? Plenary paper presented at ICAME, Apr. 2003, Guernsey, UK.

Simpson, Rita C. Defining and characterizing interactivity in academic classroom discourse. AAAL, Mar. 2003, Arlington, VA.

Simpson, Rita C. Corpus-based materials for interactive academic listening. TESOL, Mar. 2003, Baltimore, MD.

Swales, John M. A closer look at aspects of institutional speech. Workshop at the Centre for Discourse Studies research seminar, June 2003, University of Aalborg, Denmark.

Swales, John M. Evaluation in academic speech. Keynote paper presented at Evaluation in Academic Discourse, June 2003, Certosa di Pontignano, Siena, Italy.

Swales, John M. The Dissertation defense in the US. AAAL, Mar. 2003, Arlington, VA.

Swales, John M. Corpus linguistics and spoken English for academic purposes. Plenary paper presented at CIFLE 6 (6é Congrés de Llengüe per a Finalitats Especifiques), Jan. 2003, UPC-Vilanova i la Geltrú, Spain.

2002

Presentations at the 4th North American Symposium on Corpus Linguistics and Language Teaching, Nov. 2002, Indianapolis, Indiana, USA.

Briggs, Sarah & Lee, David Y.W. (Poster Presentation) Developing a Lexical Database of Academic Spoken English (LDASE) for Language Testing: Problems & Prospects.

Mendis, Dushyanthi. How do you give instructions when instructing? Evidence from a corpus of academic speech.

Simpson, Rita C. A corpus-based study comparing students' and professors' use of formulaic expressions.

Shaw, Kate & Alison Busch. Student Presentations at the University of Michigan: Argument or Show and Tell?

Swales, John M. "Any last minute thoughts on this particular search?" The occurrence of sentence-initial ellipsis (SIE) in research speech.

Other Presentations:

Simpson, Rita. C. Giving and getting advice in academic contexts: corpus-based teaching materials development. Inter-varietal Applied Corpus Studies Conference, June 2002, Limerick, Ireland.

Simpson, Rita. C. Stylistic features of academic speech: the role of formulaic expressions. BAAL, Sept. 2002, Cardiff, UK.

Simpson, Rita C., Mendis, Dushyanthi, and Komsic, Angela. A corpus-based study of idioms in academic speech. American Association for Applied Linguistics Conference, Apr. 2002, Salt Lake City, Utah.

Swales, John M. Is the university a community of practice? BAAL, Sept. 2002, Cardiff, UK.

2001

Presentations at the 3rd North American Symposium on Corpus Linguistics and Language Teaching, Mar. 2001, Boston, MA.

Mauranen, Anna. But here's a flawed argument: Socialisation into and through metadiscourse.

Pagliere, Alan. MICASE Implementation: Making the Michigan Corpus of Academic Spoken English web accessible.

Simpson, Rita C. Statistical Analysis of disciplinary Style in Transcripts of Spoken Academic English.

Swales, John M. & Burke, Amy. It's really fascinating work: Differences in evaluative adjectives across academic registers.

Other Presentations:

Briggs, Sarah & Simpson, Rita C. Using an academic corpus to evaluate the lexis of EAP tests. Language Testing Research Colloquium, St. Louis, Missouri, Feb. 2001.

2000

Presentations at the 2nd North American Symposium on Corpus Linguistics and Language Teaching, March 31- April 2 2000, Northern Arizona University, Flagstaff, AZ..

Mauranen, Anna. They're a little bit different: Observations on hedges in academic talk.

Ovens, Janine. You have no way of knowing that: A study of negation in spoken academic discourse.

Simpson, Rita C. Adverbial hedges in spoken academic language: Cross-disciplinary comparisons and teaching applications.

Other Presentations:

Mauranen, Anna. Expressing evaluation in academic speech. SESSE 5 Conference, Helsinki.

Mauranen, Anna. Pragmatized expressions in academic speech. Seventh International Pragmatics Conference, Budapest.

Simpson, Rita C. Cross-disciplinary comparisons in a corpus of spoken academic English. Teaching and Language Corpora 2000, July 19-23 2000, Graz, Austria.

1999

Mauranen, Anna. Reflexive Academic Talk: A Corpus Approach. Presentation at Dialogue Analysis VII, "Working with Dialogue", 7th IADA Conference, Apr. 1999, Birmingham, UK.

Poos, Deanna. A question of gender? Hedging in academic spoken discourse. Michigan Linguistic Society, Oct. 1999, Michigan State University, East Lansing, MI.

Simpson, Rita C., Lindemann, Stephanie & Swales, John M. First Forays into MICASE. Department of English Colloquium, Apr. 1999, Central Michigan University.

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

Appendix D
MICASE Order Form

Registration information - *Required*

The **person(s) named below** is responsible for ensuring that the terms of the **copyright and the individual/site license** are upheld:

Full Name: _____

Title: _____

Email: _____

University/Organization: _____

Please TYPE or PRINT CLEARLY all information requested below:

Please check appropriate box(es):

☐ **US$50 –** MICASE Corpus with Individual License, download version
☐ **US$180 –** MICASE Corpus with Site License (2-25 users), download version
☐ **US$10** additional charge for CD-ROM version (includes shipping & handling)

For Shipped International Orders Only

Shipping Options
○ FedEx (check FedEx website for rate)
○ UPS (check UPS website for rate)
○ DHL (check DHL website for rate)

NOTE: Unless otherwise specified, international orders will be shipped via **FedEx International Economy** where available. The ELI is NOT responsible for import taxes and/or duties. **If not prepaid, a credit card is required for all international orders.**

Select Method of Payment:

○ Check drawn on a U.S. Bank, payable to "English Language Institute" (For US Orders only)

○ International Postal Money Order, payable to "English Language Institute"

○ Credit Card: (VISA or MasterCard)

Card # _____ Expiration (MM/YY) _____/_____

Name on Card _____ Signature _____
(Sign form if submitting by Mail or Fax)

Billing Address:

First/Given Name: _____

Middle: _____

Last/Family Name: _____

Address: _____

_____

Email Address: _____

Telephone #: Country Code _____ City Code _____Phone Number _____

Shipping Address (if different from above):

First/Given Name: _____

Middle: _____

Last/Family Name: _____

Address: _____

_____

NOTE: Payment via Western Union will not be accepted. In order to ensure that this form will be processed, be sure that the card number you provide is valid. If the account is not in good standing, the process will be delayed and we will notify you. Credit card orders may be submitted by mail, fax or by using the on-line form on the MICASE web site

Feedback

We would appreciate some brief comments about how you intend to use MICASE:

- ☐ Personal research
- ☐ ESL teaching
- ☐ Linguistics teaching
- ☐ Other/comments (e.g., type of research)

_____

_____

_____

_____

Orders are shipped within 2 weeks of receipt. Rush orders (within the US only) will be shipped in 3 working days of receipt and will be assessed a $30 RUSH PROCESSING FEE in addition to shipping and handling costs. FedEx account # and/or Visa or MasterCard are required for rush orders. If no credit card is used, a purchase order number is required. Shipment will be delayed if the materials are not in stock at the time an order is placed. For questions, call (**734) 615-5606** from 9am-12pm or 1pm-4pm Eastern Standard Time, Monday-Friday.

◯ Fax (Credit Card orders only) **Fax Number: [001]-(734) 615-6586**

◯ Mail (Check, International Postal Money Order, or Credit Card orders) **Send form to:**

Office of Testing & Certification
English Language Institute
TCF Building, Suite 350
401 E. Liberty St.
Ann Arbor MI 48104-2298
USA