A Corpus-Based Study of Idioms in Academic Speech

RITA SIMPSON and DUSHYANTHI MENDIS
University of Michigan
Ann Arbor, Michigan, United States

A mastery of idioms is often equated with native speaker fluency (Fernando, 1996; Schmitt, 2000; Wray, 2000), but it is difficult for language teachers and material writers to make principled decisions about which idioms should be taught, given the vast inventory of idioms in a native speaker's repertoire. This article addresses the advantages and limitations of a corpus-based approach to researching and teaching idioms in a specific genre by drawing on a specialized corpus of 1.7 million words of academic discourse, the Michigan Corpus of Academic Spoken English. We argue that evidence from such a corpus can be quite informative for language teachers when the primary target language domain matches that of the corpus. In terms of pedagogical applications, we demonstrate the use of corpus data to construct teaching materials aimed not only at helping students learn unfamiliar idioms but also at raising their awareness of the speech contexts idioms occur in and the discourse functions they perform.

The teaching of idioms raises a number of challenging practical and research-related questions. What is an idiom? Are idioms worth teaching, and, if so, why? If idioms should be taught, which of the thousands in English should be included in any particular course or textbook, and how should they be taught? Researchers such as Fernando (1996), Wray (1999), and Schmitt (2000) equate mastery of idioms with successful language learning and native speaker fluency—a perception that many language learners share and that often translates into a desire to acquire as many idioms as possible. Moreover, the mention of the word *idiom* conjures up language that is thought to be entertaining, engaging, casual, charming, colorful, and memorable. If such perceptions are not sufficient reasons for teaching these expressions, idioms also perform many important discourse functions, further warranting their inclusion in an ESL curriculum.

The inventory of idioms in a native speaker's repertoire is indeed vast, and therefore the frequency of occurrence of any individual idiom is

relatively rare and unpredictable in any given stretch of discourse. How then can language teachers and material writers make principled decisions about which idioms are worth teaching? As Biber and Conrad (2001), Mauranen (2002), and others have demonstrated, a corpus can be used to identify the most important linguistic exemplars to teach. Although no single corpus can provide a comprehensive selection of idioms, a corpus is arguably a much better starting point than an invented list of idioms, in part because such lists are by and large entirely devoid of a coherent focus on a particular language domain—such as, for example, business or academic English. This research reports the results of a corpus-based study of the idioms of academic speech with the aim of examining the feasibility of identifying those worth teaching to students of English for academic purposes (EAP). It identifies the idioms that occur in academic speech, analyzes their functions, and offers some implications for teaching based on the corpus data.

THE STUDY OF IDIOMS

The majority of research on idioms has looked at them as a lexical phenomenon that is equally relevant across registers of English. More recently, however, attention has been directed toward idioms as a more register-specific linguistic feature.

Idioms as Formulaic Language

A number of studies consider idioms as one subcategory of the more general lexical phenomenon of formulaic language (Nattinger & DeCarrico, 1992; Moon, 1998; Wray, 1999, 2000, 2002; Wray & Perkins, 2000). A recurring theme throughout this literature is that an ability to understand and use formulaic language (including idioms) appropriately is a key to nativelike fluency. In fact, according to Fernando (1996), "No translator or language teacher can afford to ignore idioms or idiomaticity if a natural use of the target language is an aim" (p. 234). Pawley and Syder (1983) make a strong case for the daunting nature of the task learners face in figuring out which grammatically possible utterances are commonly used by native speakers—that is, which are idiomatic—and which utterances, though grammatically possible, are not nativelike. Wray (1999) supports this claim, adding that the absence of formulaic sequences in learners' speech results in unidiomatic-sounding speech. Nattinger and DeCarrico (1992) take this argument a step further and present a typology and pragmatic analysis of what they

call *lexical phrases,* along with a number of suggestions for incorporating them into the L2 curriculum.

Wray (1999, 2000, 2002) and Wray and Perkins (2000) approach the study of formulaic language primarily from a psycholinguistic perspective, arguing convincingly for a function-based account of formulaic language that acknowledges the sociopragmatic and interactional purpose of such expressions. Wray (1999) maintains that formulaic language benefits both comprehension and production, in part because such expressions appear to be stored and retrieved as holistic, unanalyzed chunks and thus contribute to economy of expression. As she puts it, "The whole point of selecting a prefabricated string is to bypass analysis" (p. 480). As stated above, because idioms are considered to be a subset of formulaic language, these claims and findings for formulaic expressions are certainly applicable to idioms as well.

Functional Descriptions of Idioms

Early studies of idioms tended to focus on formal properties such as typologies based on semantic and syntactic criteria (e.g., Makkai, 1972; Weinreich, 1969) whereas more recent work, beginning with Strässler (1982), has turned to an analysis of the pragmatic, interactional, and discourse-level features of idioms (Fernando, 1996; McCarthy, 1998; Moon, 1998). McCarthy demonstrates that idioms are highly interactive items and cannot always be identified by their formal properties. He also claims that idioms are not used randomly or without motivation and convincingly argues that they should be looked at as communicative devices rather than as mere quirks of the language (p. 146). He goes on to identify a number of socio-interactional functions for selected idioms in his corpus.

Although several in-depth studies of idioms have been conducted, to date none has examined idioms in a specialized corpus with specific pedagogical aims. Moon (1998) presents one of the most in-depth corpus-based studies of fixed expressions and idioms in English, but as she acknowledges, her corpus has certain limitations: It consists almost exclusively of written texts, and more than two thirds of those texts are of journalistic writing. McCarthy's (1998) discussion of idioms is based on a relatively large spoken corpus consisting of predominantly interactive spoken data from a number of different registers. As McCarthy points out, the distribution and function of idioms in spoken discourse are areas that need more research.

Our research, which drew on a specialized corpus of both interactive and monologic speech, therefore fills a gap not yet sufficiently addressed—

namely, the study of idioms in speech from a specific institutional context. In particular, the research questions for this study were (a) How many idioms occur in the various subregisters within academic spoken language? (b) What functions do these idioms perform?

METHOD

We began this research project to find out whether idioms occurred at all in academic speech and, if so, what conclusions could be drawn based on a corpus of under 2 million words, given that a majority of idioms have frequencies in the range of 1 token or fewer per million words (Moon, 1998). The methods of this research extended beyond that question to look at precisely how many idioms were found and what functions they served.

The Corpus

Our research was based on the Michigan Corpus of Academic Spoken English (MICASE), a specialized corpus of contemporary speech recorded at the University of Michigan between 1997 and 2001 (Simpson, Briggs, Ovens, & Swales, 2002). MICASE, which is freely available and searchable via the Web, contains 197 hours of recorded speech, totaling about 1.7 million words in 152 speech events. These speech events range from large lectures, to dissertation defenses, to one-on-one office-hour interactions and small peer-led study group sessions, and each transcript is categorized along several dimensions, including primary discourse mode and academic division. Primary discourse mode is a three-way classification referring to the level of interactivity, labeled *monologic, interactive,* or *mixed.* Academic division refers to one of four divisions defined according to the University of Michigan graduate school's classification of Departments: Humanities and Arts, Social Sciences and Education, Biological and Health Sciences, and Physical Sciences and Engineering. These two parameters were the dimensions of variation we analyzed in the quantitative part of the present study. MICASE also contains information about speaker attributes such as age, gender, and academic role (e.g., junior faculty, senior faculty, undergraduate or graduate student), but these were not taken into account in this study.

Procedures

We began by examining ESL textbooks to identify idioms, but this was not very successful. We therefore returned to a basic definition of idioms and searched the corpus.

In the exploratory phase of our research, we compiled lists of idioms from three ESL textbooks aimed at university-level learners (Madden & Rohlck, 1997; McCarthy & O'Dell, 1997; Redman & Shaw, 1999) that had been published around the same time MICASE was being compiled—that is, between 1997 and 2001. We found, however, that only about 25% of the idioms in these lists occurred in MICASE. This is partly because MICASE is a relatively small corpus, and the frequency of any given idiom in naturally occurring discourse is typically low. In addition, as we noted above, the selection criteria used by textbook authors for including idioms are somewhat unprincipled and idiosyncratic; thus it is not entirely surprising that there was little overlap between these lists and the idioms found in MICASE.

Defining Idioms

Because we needed to find idioms in the corpus that were not drawn from any list, we developed criteria for deciding what an idiom is. The relatively narrow definition we adopted resulted in a manageable database of examples to examine in detail. The most prevalent description of an idiom is *a group of words that occur in a more or less fixed phrase and whose overall meaning cannot be predicted by analyzing the meanings of its constituent parts*. Starting from the premise that an idiom is a multiword expression, we used three criteria: compositeness or fixedness, institutionalization, and semantic opacity, all of which are also noted by Fernando (1996), McCarthy (1998), and Moon (1998) in their definitions of an idiom.

Compositeness or *fixedness* means that the individual lexical units of these expressions are usually set and cannot easily be replaced or substituted for. Idioms such as *off the deep end, odds and ends,* and *making out like bandits* are all examples of such fixed expressions (attested in MICASE). *Institutionalization* refers to the conventionalization of what was initially an ad hoc, novel expression (Fernando, 1996), resulting in its currency and acceptance among the wider discourse community rather than by a small subcommunity. *Semantic opacity* indicates that the meaning of such expressions is not transparent based on the sum of their constituent parts. For example, the individual words in the idioms *tongue in cheek, on the ball,* and *put a spin on it* provide no clues to their composite meaning.

In applying the test of semantic opacity to an expression, we were

frequently reminded of McCarthy's (1998) assertion that the boundary between the opaque, idiomatic meaning of a fixed expression and its transparent, more literal meaning is often blurred. As almost all researchers of idioms have noted, defining the phenomenon to be studied is indeed problematic, and we do not claim to have solved this problem. What we have done is to arrive collaboratively at an agreed-on working definition that we refined as we proceeded.

Searching the Corpus

With these criteria in mind, we found idiom tokens by intensively reading a selection of the transcripts in the corpus and then searching the entire corpus for those tokens or variations thereof (e.g., *twisted/twisting my/your/his/her arm, arm twisting*) using a concordance program (WordSmith Tools; Scott, 1996) in order to obtain overall frequency counts for each idiom. Transcripts were randomly selected from each of the major academic division and speech event categories, and each one was read through in its entirety. Although we do not claim to have uncovered all the idioms in MICASE, we have read through at least half of the transcripts in this manner and are confident that we have a representative and relatively complete inventory.

To compile our master list of idioms, we used the three criteria supported by the collective intuition of three raters to eliminate some expressions. For example, expressions that may be considered metaphorical but are not idioms (e.g., *a sad showing*) were eliminated, as were phrasal verbs (e.g., *catch on, bounce back, jump into*) because these expressions constitute a separate subcategory often considered to be simply idiomatic verb phrases. In addition, we eliminated certain binomial expressions that we considered to be semantically transparent—for example, *pick and choose, little by little, trial and error*—but we included such binomial expressions as *odds and ends, nuts and bolts,* and *crash and burn* because we felt that they satisfied the criterion of semantic opacity. This master list of idioms was entered into a database (created with Microsoft Access) that was linked to the existing database of all the speech events in the corpus. The linked database allowed us to compile various sorted lists of all the idioms and the transcripts they occurred in, which we used to obtain the frequency distributions. Finally, we used a discourse-analytic approach to identify and illustrate the primary pragmatic functions of a selection of idioms in the corpus.

FINDINGS

Freqencies and Distribution of Idioms in MICASE

Two main questions guided the quantitative analysis. First, are there more idioms overall in the interactive transcripts of the corpus than in the monologic ones? Idioms are often assumed to be more prevalent in informal and interactive communication; certainly, a glance at many of the popular ESL textbooks purporting to teach idioms lends support to this belief. Second, do the overall frequencies of idioms show any noticeably differing trends across the four academic divisions? Preliminary forays into the corpus led us to suspect that we would find more idioms in the humanities and social science divisions than in the hard sciences, perhaps reflecting the stereotypical expectation of encountering colorful and rhetorically sophisticated speech in the arts and humanities and technical speech in the hard sciences and engineering. Both of these assumptions appear to be unfounded, according to our results.

In all, we found 238 idiom types (unique idioms), with 562 tokens in the corpus, that met our criteria (see Table 1, which shows four frequency ranges and the number of types in each range). Of these, 123, or over half of all types, occurred only once, and only 23, or 10%, occurred more than four times in the corpus. As for the frequency of occurrence across the four academic divisions and primary discourse modes, the results of our research show a remarkably patternless distribution (see Table 2). Expressed in terms of either instances (i.e., omitting repetitions of the same idiom in a given transcript) or total tokens (i.e., counting all repetitions) per 100,000 words, neither the humanities nor the hard sciences show striking differences, and, similarly, idiom frequencies were only slightly higher in the monologic than in the interactive speech events.

TABLE 1
Idiom Types in MICASE by Frequency of Occurrence

Range of occurrences		
Raw frequency	Frequency per million words (approximate)	No. of idiom types in range
10–17	5.8–10.0	8
5–9	3.0–5.3	15
2–4	1.2–2.4	92
1	0.6	123

Note. Total idiom types = 238; total idiom tokens = 562.

TABLE 2
Frequency Distributions of Idioms in MICASE by Primary Discourse Mode and Academic Division

Category	Transcripts	Words	Per 100,000 words: Idiom instances[a]	Per 100,000 words: Idiom tokens[b]	Total: Idiom instances[a]	Total: Idiom tokens[b]
Primary discourse mode						
Monologic/panel	70	688,875	30	37	203	256
Interactive	57	714,906	25	32	180	227
Mixed	25	284,323	21	25	60	72
Academic division						
Inter-/nondepartmental	13	156,171	37	45	58	70
Social Sciences & Education	33	394,780	31	41	123	162
Physical Sciences & Engineering	36	358,729	24	32	85	113
Humanities & Arts	37	447,487	24	31	110	137
Biological & Health Sciences	33	330,937	20	22	67	73

[a]One token of an idiom occurring in a transcript, not counting multiple tokens. [b]All occurrences of idioms.

As further evidence that idioms are distributed across a wide range of academic speech events, even after a not quite exhaustive search, as described above, we found that 11 transcripts of a variety of speech events contained 10 or more idioms. The speech events included in these transcripts were as diverse as an organic chemistry study group, a meeting between a graduate student and his adviser, a class on ethics issues in journalism, and a large public lecture by a recent Nobel laureate in physics.

In addition to the two primary, corpus-internal comparisons, another general question motivating the quantitative analysis was whether or not idioms would turn out to be more or less frequent in academic speech than in any other spoken genre. We speculated that idioms might prove to be rare in academic speech compared with corpora of general conversation, but no studies have attempted to quantify the number of idioms in any genre. Moon (1998) compared the frequencies of 24 idioms in the entire Bank of English (a continually growing corpus, currently containing over 400 million words of spoken and written English; see Collins Cobuild, n.d.) with their frequencies in the 20-million-word subcorpus of conversation and found only five of the items to be more common in conversation. She sums up this finding by speculating that "people may be impressionistically overreporting high

incidences of idioms in conversation" (pp. 72–73). In light of these findings, we would surmise that idioms are neither rare nor particularly frequent in academic speech or any of its subgenres. The overall frequency of idiom tokens in MICASE, using our rather narrow criteria, is about 330 total tokens per million words, or 260 per million not counting repetitions within the same speech event. Thus taken as a whole, these items constitute a not-insignificant feature of the lexical landscape of academic speech. At the same time, quantitative comparisons of idioms are difficult to assess because of the relative infrequency of any given idiom combined with the difficulty of ensuring that other researchers have followed similar criteria in deciding what counts as an idiom.

Pragmatic Functions

As mentioned above, McCarthy (1998) strongly supports a discourse-oriented approach to the study of idioms in spoken language, which he points out has not been thoroughly researched for the occurrence and functions of idioms. Accordingly, we discuss the most salient discourse functions associated with certain idioms in MICASE that best embody these functions. Of the functions identified here, several (paraphrase, emphasis, and metalanguage) are particularly relevant in the context of academic discourse. However, by identifying one particular discourse function for an idiom, we do not mean to overlook the fact that a single idiomatic expression often performs more than one function. Moon (1998) refers to this phenomenon as *cross-functioning,* or the use of an expression in functions other than their primary or most obvious one in the discourse. Many of the idioms we discuss below exhibit such cross-functioning.

Evaluation

Like McCarthy (1998), who illustrates evaluative uses of idioms in conversational interactions, we found idioms used for these purposes in academic discourse. Two examples from MICASE that illustrate this function are *out of whack* and *threw her for a loop.* Example 1, while providing an evaluation, is also an instance of what McCarthy refers to as the "observation plus comment function" (p. 133), in which evaluative comments are preceded by a factual observation. Example 2 illustrates the observation made by Strässler (1982) and supported by McCarthy that when speakers use an idiom for evaluation, it is much more likely to occur in a third-person context, especially if it represents a threat to face.

1. the paintings didn't seem to use the conventional skills of perspective the perspective is all *out of whack* here. um and that the paint itself is kind of slopped onto the canvas with these heavy dark art- outlines, without the. . . . (Visual Sources lecture)[1]

2. they would just be like, what are you talking about? you know just, suck it up deal with it hang with it you know and, she_ it really *threw her for a loop* and she actually um, she came_ she, was in the math department she then chose to transfer to the statistics department (Women in Science conference panel)

We find, however, that under certain circumstances, face-threatening idioms can be used in a second-person context. Example 3 is taken from a study-group interaction between two students. The two speakers are peers and obviously friends, as the interaction is tempered with both sarcasm and humor, two important factors that mitigate the face threat implicit in the idiom.

3. S2: what were you saying?
   S3: I was mumbling something to myself, I guess.
   S2: oh I'm sorry that I interrupted your conversation with yourself.
   S3: I thought I was talking to you but I guess not.
   S2: <Laugh> you were just mean to me again. <S3 Laugh> oh you've got this *evil side that rears its ugly head.* . . . (math study group)

Description

We also found idioms used for description, a function that often overlaps with that of evaluation, representing an instance of cross-functioning, discussed above. Evaluative uses are often also descriptive (as in Example 1), but description does not always entail evaluation. Examples of idioms used in this manner in MICASE are *hand in hand, run of the mill,* and *out of whack.*

4. the Roman empire was not acquired for, economic reasons. this is very different, from European empires, uh in the eighteenth and nineteenth centuries. uh where both national pride and, uh notions of economic expansion go *hand in hand.* (lecture on sports in ancient Rome)

[1] Transcription conventions are as follows:

< >	contextual comment
[S1]	speaker identification
[SU]	unidentified speaker
.	words omitted
(xx)	unclear speech
word_	truncated speech

5. the reason why we like really good, premium ice creams, is because they have high fat content, that's the main thing that separates a premium ice cream from your *run-of-the-mill* ice cream. (Introduction to Psychology lecture)

6. if your thyroid is severely *out of whack,* it has all the exact same symptoms of depression. (Introduction to Psychopathology lecture)

The primary function of each of these expressions is descriptive, and in Examples 4 and 5 the idiom is used to highlight a contrast. Example 6 is an instance of a speaker using an idiom for economy of expression—the unanalyzable chunk *out of whack* here substitutes for *not functioning properly* or another more detailed, literal description.

Paraphrase

Another function closely related to description is that of providing a paraphrase or gloss of the discourse content. This function is particularly well suited to academic spoken discourse, given its heavy use of explication. Examples from MICASE illustrating this use are *put up a stink, no mean feat,* and *a dime a dozen*, which also illustrate McCarthy's (1998) observation-plus-comment function. Such paraphrasing uses of idioms often have the effect of reducing the distance between the speaker and listener, through the juxtaposition of either a formal academic word or a longer literal phrase with a more colloquial expression (as in Example 7). Moon (1998) discusses a similar interactional function in her analysis of idioms and other fixed expressions. The use of a casual, almost slang expression could indicate an attempt on the part of a speaker to reduce the formality and highly transactional nature of academic discourse.

7. so women knew that uh if they were labeled noncompliant if they were you know *put up a stink* about where they delivered their children, this might have a negative impact on health care for their whole family, ... (Medical Anthropology lecture)

8. at the beginning of airlifting women out it was not a very easy thing to do getting those women out of there to deliver them in hospitals was *no mean feat.* (Medical Anthropology lecture)

9. we conclude that uh, {[SU] we win.} these these uh, you know percentages of of favoring Bush are a *dime a dozen* they they they, they occur so often, that what's more important is is, a key decision of a reversal that Bush made. . . . (Ethics Issues in Journalism discussion/lecture)

Emphasis

Speakers also use idioms to emphasize content or reinforce an explanation, two functions that complement the goals of academic discourse particularly well. In such cases, there is a tendency for the speaker to repeat the idiom, often with truncation or creative variations. One such instance is the idiom *carrot and stick,* used by a lecturer six times in the span of about 5 minutes to explain the concept of rewards and punishments in the context of international relations. Another such case is Example 10, where *the kitchen sink* is used three times with variations to refer to a statistical concept. *Put the heat on,* discussed later, serves a similar function, where multiple speakers use variants of the idiom to reiterate or emphasize the same point. In fact, idioms are often associated with a certain amount of hyperbole, which is closely related to emphasis. Examples 8 and 9 above, identified as paraphrasing uses, both add a degree of emphasis as well.

10. mostly in, doing data analysis we're interested in posing particular questions that're interesting to us, and we're not interested in explaining the total variance in our outcome by *throwing everything in but the kitchen sink.* so, by looking only at R-square of how good is our model that would be the kind of what I would call the *kitchen sink* model, and in education data even the *kitchen sink* model is not gonna do you very well, the_ unless you have, unless your outcome is . . . . (Statistics in Social Science lecture)

Collaboration

Idioms can also be used to create collaborative discourse and establish a sense of solidarity within a group of speakers. Like emphasis, collaborative uses of idioms are often achieved through repetition. McCarthy (1998) found idioms performing the same function in his conversational data when participants express shared views and ideas. A striking example of this from MICASE is the way multiple speakers use *put the heat on* in discussing ethics in journalism. In fact, in this extended excerpt (Example 11), the discussion at its most animated point actually revolves around the idiom itself. The example also shows how certain idioms allow variations in form. The base form of the idiom, *put the heat on,* produces at least six creative or syntactic variants from different speakers. Thus we have *under some heat, puts some heat on, put heat on, putting heat on themselves, heat put on them,* and *put more heat on them.*

11. S9: just I, I uh tend to think why is it anybody's business, to see these photos? I mean, the guy is dead, (. . .) and if four drivers have died

in the last 9 months, I'm sure NASCAR will take some steps to correct that and it, doesn't require the involvement of journalists.

S1: are they more likely to take steps if the press *puts some heat on them*?

SS: mhm yeah

S2: the press has already *put heat on 'em* they're *putting heat on themselves*. I don't think it's necessary.

S1: well they're obviously not having enough *heat put on them* because it keeps happening. {S2: I think (xx)} I mean for_ at least it's possible. yes?

S10: showing pictures of a dead body is not gonna *put the heat on 'em* that's necessary to make the changes though, I mean it's gonna be the reporting on the incident, you know it's th- one photo of a dead_ I mean I i just think that it's, definitely an invasion of privacy, that isn't necessary.

(several minutes/turns later:)

S1: yeah, there's the danger it will get published, number one number two it's not clear to me I don't know how you feel but it's not clear to me that, having someone look at this photo's gonna make any difference as you said, whether it's gonna *put more heat on* NASCAR I don't, it doesn't seem clear. (Ethics Issues in Journalism discussion/lecture)

Metalanguage

In academic discourse, the use of idioms in the metalanguage is particularly interesting. In her analysis of idioms, Fernando (1996), among others, observes that the academic register is a highly complex one and requires "distinctive organization and signposting devices which function as creators of coherence and intelligibility" (pp. 232–233). Moon (1998) mentions a similar organizational function for fixed expressions and idioms, in which they organize texts by signaling logical connections between, for example, propositions and summaries. Some of the idioms in our data that clearly function as such signals or signposting devices are *go off on a tangent, on that note, cut to the chase,* and *train of thought*. In Example 12 the speaker uses the idiom to signal that the commentary is about to move in a different though not entirely irrelevant direction. In Example 13 the idiom signals a change of focus, with the added implication that the discussion is drawing to a close. In Example 14 the idiom functions as an organizational device to create coherence in the discourse, and in Example 15 the idiom is a filler commenting on the apparent lack of coherence. In each of these examples, the meaning of the idiom is much more closely tied to the function than in the previous examples.

12. and the Basques by the way have been some in the history of Spain, if I may just uh, *go off on a tangent* the Basques have been, in Spanish history and Spanish religious history some of the most fervent Catholics. (Historical Linguistics lecture)

13. well *on that note,* have another cup of coffee on your way out and let's thank John and Ivette for a really nice (xx) (Ecological Agriculture colloquium)

14. what to do? what to do? lemme do it this way. you will never stand for this derivation I know it. <Laugh> not today. the weather's too nice, so I'm gonna *cut to the chase* and I might come back and fill in some details on Monday, (Chemical Engineering lecture)

15. um, uh... the issue, comes down to, lost my thou- I almost *lost my train of thought* here. um, how . . . . (graduate education advising)

PEDAGOGICAL APPLICATIONS

The discovery of a significant number of idioms in a corpus of academic speech and, more importantly, the evidence that they perform a variety of important pragmatic functions provides the rationale for including them in an EAP curriculum. We therefore outline our general approach to the teaching of idioms and offer some suggestions for incorporating corpus data on idioms into classroom materials by introducing some corpus-based pedagogical materials that we developed for our own teaching. We then discuss some of the challenges the learner faces in attempting to acquire English idioms.

Like Wray (2000), in reference to formulaic sequences more generally, we advocate striking a balance between a holistic approach that focuses on learning idioms as chunks, that is, paying attention only to their composite meaning, and an analytical approach that teaches the meaning of an idiom by explaining the meaning of its constituent parts. In L2 teaching, some idioms lend themselves better to the latter approach even though most native speakers store them as holistic chunks. For example, students are much more likely to understand and remember the idiom *a drop in the bucket* if they know what *drop* and *bucket* mean, as the metaphor is rather transparent even if speakers do not commonly associate the literal meaning with its composite, metaphorical meaning. Similarly transparent idioms include, for example, *on the fringe, shift gears,* and *making out like bandits*. In contrast to these are idioms like *run of the mill, cut to the chase, one fell swoop,* or *on that note,* none of which embodies transparent conceptual metaphors for which knowledge of the lexical components or origins of the expressions is likely to contribute to

learning and remembering their forms and meanings. The point here is that when the conceptual metaphor of a given idiom has a relatively high degree of transparency, it may well be advantageous, as Wray (2000) puts it, to "[legitimize] the classroom learners' inherent desire and ability to analyse" (p. 484). When the holistic meaning of the idiom is too far removed from the domains of the conceptual metaphor, however, it makes much more sense to resist the desire to analyze and encourage students to learn the idiom purely as a chunk. The teacher's task, then, is first to convince students to learn some—if not most—idioms as chunks and not attempt a constituent analysis, which becomes in many cases a futile exercise with little payoff in terms of learning outcomes. Second, as teachers, we must also recognize that some metaphorical idioms lend themselves to a constituent analysis and that we would do students a disservice to insist that they learn all idioms as unanalyzed chunks simply because this is the way native speakers seem to learn and process them.

Another strategy we advocate is to incorporate the notion of discourse function into the teaching of idioms. As mentioned above, some idioms are in fact used in quite predictable ways, so that the pragmatic function is integral to the meaning. For instance, *on that note* and *shift gears* are typically used at a discourse boundary, signaling the end of one topic or episode and a transition to another. Similarly, the idiom *lost my train of thought* is almost always used in a situation of disfluency. Other idioms, as previously noted, may be used for a variety of pragmatic functions, but even though the correspondence between function and meaning is not as strong, teaching idioms from a function-based perspective still offers significant advantages. Students should be made aware of the discourse functions (e.g., evaluation, emphasis, paraphrase) idioms are associated with and likewise should be taught to identify them.

Corpus-Based Teaching Materials

According to McCarthy (1998), idioms are highly interactive, engaging both the speaker and the listener, and are therefore best studied in context, yet they tend to be taken out of their contexts and taught as disembodied items. Moreover, if a context is provided, it tends to be an imagined and contrived one. Using real speech samples from contexts that learners will be exposed to has distinct advantages over using conventional methods for teaching idioms. Also, although the highly interactive nature of idiomatic expressions calls for a discourse-oriented pedagogical approach, for practical reasons such an approach may not be possible in every classroom. To compensate for the difficulty of constructing the appropriate interactional climate for the teaching of

idioms, McCarthy proposes the raising of students' awareness of idiom usage as a first step.

At several workshops on idioms we have conducted at our university's English Language Institute, students responded positively to an approach that began with consciousness raising and moved on to the introduction of idioms in authentic discourse contexts. First, we briefly discussed the nature of an idiom and had students identify opaque idiomatic expressions in excerpts from transcripts of spoken discourse. Next, we gave them the contexts of selected idioms from MICASE and helped them identify the lexical and semantic cues in a speaker's utterance that would be helpful in understanding the meaning of the idiom. In this exercise, we found that contextual extracts rich in discourse markers signaling, for example, similarity, contrast, and sequence lend themselves particularly well to teaching. Similarly, a high frequency of content words provides learners with additional contextual clues that help in processing the meaning of an idiom. The ideal context for pedagogical purposes, however, is one in which the idiom is preceded or followed by a paraphrase or gloss, such as in Examples 7–9.

In terms of pedagogical materials, some students responded well to a multiple-choice exercise designed to summarize the meaning of an unfamiliar idiom (see Exercise 1 in the Appendix). However, excerpts from the corpus that provided a context for each idiom proved to be the most popular (Exercise 2 in the Appendix). In selecting such excerpts for teaching, teachers must take care to choose excerpts that are rich in contextual clues, as described above. We also used audio recordings of selected excerpts from the corpus to facilitate listening comprehension and to attune students to how idiomatic expressions sound in native speaker interactions. The excerpt in Example 11 illustrating the use of *put the heat on* was particularly effective as a listening exercise because the same idiom recurs frequently in an extended stretch of discourse.

Based on our experience, we would not recommend the use of single, unedited concordance lines of spoken text as teaching materials. Such data were not as compelling and seemed to pose comprehension difficulties for the students. These difficulties can be avoided if the examples are cleaned up slightly and presented as individual excerpts with more than a single line of context.

Additional teaching suggestions arising from this study include comparing and discussing perceived differences in the communicative effects of using an idiom versus an alternative literal expression (with the caveat that such differences can sometimes be difficult to ascertain). For example, possible communicative effects as referred to in the section on pragmatic functions include exaggeration, informality, and rhetorical flair. Again, the examples illustrating paraphrase, in which the literal expression occurs with the idiomatic expression, are particularly useful

in this regard. Another valuable exercise would be one that showed the same idiom used for different discursive functions, to reinforce the idea that idioms are not necessarily bound to a single function. Finally, some general observations about the frequency of idioms may be worthwhile, such as the fact that they are not limited to strictly informal contexts, as illustrated by the Nobel laureate physics lecture mentioned above.

Another useful resource that we culled from this research was a list of 20 idioms (see Table 3) that lend themselves particularly well to an academic context, based primarily on their semantic content but also partly on frequency. Although some of these idioms have a low frequency in our corpus, we believe they provide a useful starting point in compiling a list of idioms for a future EAP curriculum. Finally, we present here a list of the most frequent idioms in MICASE: 32 idiom types that occurred four or more times in the corpus (Table 4). Whereas this list is not an authoritative or definitive list of the most common idioms in academic speech, given the stated limitations of a 1.7-million-word corpus, it is nevertheless a worthwhile resource. Some of these high-frequency idioms were used by only one speaker or in one speech event, or appear in contexts that do not lend themselves particularly well to teaching, but many are prime candidates for inclusion in an EAP or other advanced ESL curriculum.

Difficulties for Learners

One final point to keep in mind is that idioms in authentic discourse do not always occur in their canonical forms. Such variations could pose difficulties in comprehension when learners encounter idioms as actually

TABLE 3
Particularly Useful Idioms for English for Academic Purposes Curricula

Idiom	Total tokens in MICASE	Idiom	Total tokens in MICASE
bottom line	17	hand in hand	8
the big picture	7	hand-waving	2
carrot and stick	7	in a nutshell	3
chicken-and-egg question	1	ivory tower	3
come into play	16	litmus test	1
draw a line between	7	on the same page	4
get a grasp of	1	play devil's advocate	3
get a handle on	4	shift gears	1
get to the bottom of things	1	split hairs	4
go off on a tangent	3	thinking on my feet	1

TABLE 4
Idioms Occurring Four or More Times in MICASE

Idiom	Total tokens	Idiom	Total tokens
1. bottom line	17	17. out the door	6
2. the big picture	16	18. rule(s) of thumb	6
3. come into play	14	19. take (something) at face value	6
4. what the hell	12	20. beat to death	5
5. down the line	11	21. put the heat on	5
6. what the heck	10	22. a ballpark idea/guess	4
7. flip a coin; flip side of a/the same coin	10	23. come out of the closet	4
8. on (the right) track	9	24. full-fledged	4
9. knee-jerk	8	25. get a handle on	4
10. hand in hand	8	26. goes to show	4
11. right (straight) off the bat	7	27. nitty-gritty	4
12. carrot(s) and stick(s)	7	28. on the same page	4
13. draw a/the line (between)	7	29. ring a bell	4
14. on target	7	30. split hairs	4
15. thumbs up	7	31. take (make) a stab at it	4
16. fall in love	6	32. take my/someone's word for it	4

used by fluent speakers. A number of idioms are subject to truncation. Examples from MICASE include *haven't the foggiest,* where the word *idea* was left out; *rearing its head,* where the word *ugly* was left out; and simply *carrot* (used to refer to *carrot and stick*). Speakers subject idioms to what Barlow (2000) refers to as *creative blending*. We found *flip side of the same coin*, with the insertion of the word *same* (presumably for emphasis), and an even more creative derivational extension, *coin flipping*. We also found *arm twisting,* from *twist someone's arm* and *one nit to pick,* from *nitpicky. The kitchen sink model* and *under some heat,* referred to above, are two more examples of this process. McCarthy (1998), commenting on this phenomenon, observes that speakers use idioms creatively "by a process of 'unpacking' them into their literal elements and exploiting these" (p. 137).

We also found *performance variations,* a term we use to describe idiomatic expressions that are not instances of conscious creative manipulation or exploitation as described above but rather appear to be spoonerisms or the accidental substitution of a different verb, noun, or pronoun and, in at least one case, the use of the idiom for a meaning entirely different from its commonly understood one. Some examples are *walking through a landmine* instead of *walking through a minefield, side in your thorns* instead of *a thorn in your side,* and *pick up where we took off* instead of *pick up where we left off.* The idiom used with a different meaning was *take a stab at,* used to mean *criticize* instead of its usual *attempt*

to do something. Although none of these variants is likely to cause comprehension difficulties to a fluent or native speaker, a learner who uses a reference guide to idioms or a textbook with only the "pure" forms of an idiom could indeed encounter some difficulties with these forms. By using authentic examples and including naturally occurring variants, students can be alerted to the fact that despite the generally fixed nature of idioms, creative and unpredictable uses occur.

Another set of interesting but potentially problematic idioms are those that rely on and assume a specific cultural schema for interpretation. From the MICASE data, we identified *cry wolf, in bed with, wearing the pants, the same ballpark,* and *revolving door* as examples of such idioms. Each of these expressions is bound to a specific social or cultural context that may or may not be part of a learner's schema. For example, *wearing the pants* alludes to the person in a household who has the ultimate authority, assumed at one time to be the man, who was traditionally the one who would wear pants; by extension, the idiom refers to whoever has the implied authority in a particular situation. To understand the meaning of the idiom *in bed with* first necessitates associating the literal phrase with the concept of intimacy and then making the further analogy of two parties having a secret affair to two companies or organizations granting each other illicit favors or privileges. These idioms are among those that would best be taught through explanation and analysis rather than merely as unanalyzable chunks.

CONCLUSION

Our research has shown, for one, that idioms occur in academic speech and are not as rare a phenomenon as they might appear when taken as a whole. Secondly, the distribution of idioms in the subgenres of academic speech seems not to be predictable on the basis of categories of either level of interactiveness or academic division. Rather, we would conclude that the use of idioms seems to be a feature more of individual speakers' idiolects than of any linguistic or content-related categories. Some speakers in our corpus used idioms quite frequently whereas others rarely did, regardless of their socio-interactional roles. In terms of a functional discourse grammar of academic speech, the role of idioms cannot be ignored, as they clearly fulfill several important functions particularly relevant to the unique discourse features of this speech genre. Finally, we hope we have shown that a specialized corpus such as MICASE provides a rich resource for teaching materials that allow teachers not only to use authentic, attested examples of idioms in context but also to consider larger issues of discourse and sociopragmatics. This type of resource should go a long way toward relieving teachers of

the need to create contrived contexts for idioms and teach them as disembodied items.

Two important methodological advantages ensue when a corpus can be consulted for examples of idiom usage. First, the idioms can be presented in authentic contexts rather than in the contrived ones often found in textbooks or thought up by teachers. A second and closely related methodological benefit of using a corpus is that idioms can then be taught from a discourse perspective rather than as isolated lexical items, with attention not only to their immediate context but also to their sociopragmatic and interactional features. Both points are important because another challenge of learning idioms is developing an awareness of when it is appropriate to use a particular idiom. A specialized corpus that directly reflects the students' universe of discourse is particularly valuable in this regard, as it provides not only attested examples of idioms in use but examples embedded in contexts that learners will find familiar and relevant.

ACKNOWLEDGMENTS

An earlier version of this article was presented at the 2002 conference of the American Association for Applied Linguistics in Salt Lake City. We gratefully acknowledge the contributions of Angela Komsic, our coauthor on that paper. We also thank the two anonymous reviewers and the editor of this special issue for their insightful suggestions for revising this article, and our colleagues at the English Language Institute, who provided helpful comments at different stages of our research.

THE AUTHORS

Rita Simpson is a research associate at the English Language Institute of the University of Michigan. She is currently project director for MICASE and served as the founding project manager from 1997 until 2002. Her research interests include discourse analysis and pragmatics of academic speech as well as various topics in corpus linguistics.

Dushyanthi Mendis is a senior lecturer at the Department of English at the University of Colombo. She is currently a doctoral student at the University of Michigan's Department of Linguistics, researching the use of metaphor in academic speech.

REFERENCES

Barlow, M. (2000). Usage, blends, and grammar. In M. Barlow & S. Kemmer (Eds.), *Usage-based models of language* (pp. 315–345). Stanford, CA: CSLI.

Biber, D., & Conrad, S. (2001). Quantitative corpus-based research: Much more than bean counting. *TESOL Quarterly, 35,* 331–336.

Collins Cobuild. (n.d.). *Corpus concordance sampler.* Retrieved May 30, 2003, from http://titania.cobuild.collins.co.uk/form.html

Fernando, C. (1996). *Idioms and idiomaticity.* Oxford: Oxford University Press.
Makkai, A. (1972). *Idiom structure in English.* The Hague: Mouton.
Madden, C., & Rohlck, T. (1997). *Discussion and interaction in the academic community.* Ann Arbor: University of Michigan Press.
Mauranen, A. (in press). Spoken corpus for an ordinary learner. In J. Sinclair (Ed.), *Corpus linguistics and language learning.* Amsterdam: Benjamins.
McCarthy, M. (1998). *Spoken language and applied linguistics.* Cambridge: Cambridge University Press.
McCarthy, M., & O'Dell, F. (1997). *Vocabulary in use: Upper intermediate.* Cambridge: Cambridge University Press.
Moon, R. (1998). *Fixed expressions and idioms in English.* Oxford: Clarendon Press.
Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching.* Oxford: Oxford University Press.
Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–226). New York: Longman.
Redman, S., & Shaw, E. (1999). *Vocabulary in use: Intermediate.* Cambridge: Cambridge University Press.
Schmitt, N. (2000). *Vocabulary in language teaching.* Cambridge: Cambridge University Press.
Scott, M. (1996). WordSmith Tools (Version 3.0) [Computer software]. Oxford: Oxford University Press. (Available from http://www.liv.ac.uk/~ms2928/)
Simpson, R. C., Briggs, S. L., Ovens, J., & Swales, J. M. (2002). *The Michigan corpus of academic spoken English.* Ann Arbor: The Regents of the University of Michigan. Retrieved June 4, 2003, from http://www.hti.umich.edu/m/micase/
Strässler, J. (1982). *Idioms in English: A pragmatic analysis.* Tübingen, Germany: Gunter Narr.
Weinreich, U. (1969). Problems in the analysis of idioms. In J. Puhvel (Ed.), *The substance and structure of language* (pp. 23–81). Berkeley: University of California Press.
Wray, A. (1999). Formulaic language in learners and native speakers. *Language Teaching, 32,* 213–231.
Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics, 21,* 463–489.
Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge: Cambridge University Press.
Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: an integrated model. *Language and Communication, 20,* 1–28.

APPENDIX

Corpus-Based Exercises

Exercise 1: Explaining the Meaning of Idioms

Circle the answer that best explains the meaning of the idiom.

1. *take a stab at*
   a) try to do
   b) criticize
   c) fail at
   d) betray

2. *on target*
   a) fixed as an absolute
   b) completely accurate
   c) not moving
   d) busy at work
3. *shift gears*
   a) wait for a few minutes
   b) end the discussion
   c) move to a different topic
   d) start an argument
4. *tune someone out*
   a) ignore them
   b) go somewhere with them
   c) praise them
   d) misunderstand them
5. *keep tabs on*
   a) agree with something
   b) continue at the same pace
   c) observe or record carefully
   d) keep a secret
6. *odds and ends*
   a) the final events
   b) strange events
   c) harsh words
   d) various small items
7. *garden variety*
   a) new and exciting
   b) common type
   c) forbidden or illegal
   d) colorful
8. *take the plunge*
   a) commit to something important
   b) make a large profit
   c) set unrealistic goals
   d) appear out of nowhere
9. *off the wall*
   a) a waste of time
   b) dangerous or risky
   c) odd or unexpected
   d) very important
10. *take someone to task*
    a) scold or criticize them harshly
    b) talk to them privately
    c) buy them lunch
    d) encourage them to succeed

Exercise 2: Idioms in Context

Now read the following excerpts from MICASE, and see if you can figure out the meaning of the idiom from its context.

1. there were some definitions in the reading anyone, wanna *take a stab at it*? mhm?
2. estimator three has the lowest overall variability, even though it might not be *on target*.
3. anything else about The American that we wanna *touch on*? well let me *shift gears* for a few minutes then to prep you for what we'll do next week.

4. my boss talks to himself all the time and I've learned to *tune him out* when he's doing that so I don't even hear what he says.
5. how about the roles of the other people in the group? 'cuz you've been kind of the minder in a way *keeping tabs on things* and keeping things going.
6. I've got a few pieces of business and *odds and ends* I have to deal with and I want to say a little bit about how things are going with the projects so I'd like to reconvene if we could at half past nine.
7. next slide shows we do, in addition, we do have your sort of *garden variety* of clear cutting, so there is some extensive forest cutting going on, although the present time, that's sort of stagnated because of . . . .
8. for the time of the post-doc, you know, others will expect you to function something like an assistant professor. so again it just depends on the institution but it's a wonderful opportunity for you to move to that next step before you *take the plunge* of getting on a tenure-track line.
9. wow, that's an interesting question um, you know, I mean the first thing that comes to my mind is this may be a little *off the wall* but the first thing that comes to mind is this has done a lot for me as a musician but I haven't had the time to actually be a musician, so I'm looking forward to how this will transpire in my practice and in my playing. this, really doesn't have much to do with, the academic part but the musician was sort of on hold for a year . . .
10. the poem basically carries that, you know that old-fashioned, reactionary idea that the primary purpose for women in life is to bear children and that's what this poem is all about. Diane Di Prima, another San Francisco poet, *took him to task* on the subject, and you'll find her poem on page three sixty-one. (. . .) so it's good to know that she was she was ready to take him to task, on that subject. that'd be kind of an interesting thing for you to study if you're talking about gender, relations and gender ideas in the Beat Generation would be to contrast, that poem by Gary Snyder with that poem, uh, by Diane Di Prima.
11. that fifty-fifty means in terms of the money, spent. half the money is spent on treatment, half the money is spent on law enforcement. maybe the half that's spent on treatment goes a lot farther. you get more *bang for your buck* in that fifty than you get in the other fifty.