6

An Investigation of Relationships between Spoken and Written Academic English: Lexical Bundles in the AWL and in MICASE

LUCY PICKERING AND PATRICIA BYRD

Abstract

The relationship between the language that students read in preparation for their classes and then hear in the classroom is complex. Not only does spoken academic language prioritize a different set of expressions than its written counterpart, even technical terms must be understood as they are represented by acoustic characteristics in the speech stream rather than as written forms. In this chapter, we use two corpora of academic English, the Michigan Corpus of Spoken Academic English (Simpson-Vlach, & Leicher 2006) and the Academic Word List (Coxhead, 2000) to examine some of these differences, specifically as they apply to four-word lexical bundles. In a comparison of the most frequently occurring bundles in each corpus, we propose that crucial differences in the structural and functional characteristics of bundles in each register suggest that one system of categorization is not sufficiently explanatory for both written and spoken academic discourse. A lack of recognition of the different discourse cues that prevail in a spoken versus written modality (i.e., intonation structure) can obscure crucial communicative differences between spoken and written realizations of the same (or similar) phrases and the variety of pedagogical purposes that they may perform. We address a number of important pedagogical implications that arise from our analysis and suggest that students will benefit from specific instruction on how to mediate between parallel written and spoken forms.

Connecting What Students Read to What They Hear in Class

In a reading assignment from a college or university textbook, students are likely to find the word *percent* used in sentences such as the following from the corpus created for the Academic Word List (AWL) (Coxhead, 2000)[1]:

> Rural dwellers comprised twenty percent of the population in 1954, dropping to sixteen percent in 1967. . . .

When students are in class listening to a university instructor, the same word is more likely to be heard in a statement like the following example from the Michigan Corpus of Academic Spoken English (MICASE):

> . . . however this compound uh binds to the extent of ninety-nine-point-eight percent, uh ninety-nine-point-five percent . . .

While the core meaning of *percent* remains essentially the same in both contexts, the use of the word in writing and speaking involves substantial differences that could be confusing for students as they attempt to connect their reading with what they hear in class.

The spoken data, for example, includes a number of disfluencies; in addition, *percent* is unstressed or deaccented by the speaker in these environments as it is assumed that the listener understands that this word rather than any other will be selected based on the context of the interaction (Brazil, 1997). In academic tasks in which students are working with their peers, the discourse context may become even more complex for the NNS listener. The following example from the MICASE corpus is replete with interactional discourse markers such as *like* or *you know what I mean* that make it increasingly difficult to follow the progression of the idea:

> Did you say that this, like from your research like, it was credited to Solomon? because like, when he wrote a lot of this stuff it was, after, like his, like he had turned away from God and come back, and it seems like... i don't know i just had this funny thought that, like, like if he had that kind of experience, like, if it was a man writing this it would be more likely to be about, like the actual, like

[1] We are grateful to Averil Coxhead for providing data from the AWL corpus in a format that was useable for our research and at the same time maintained the copyright security that she built into her corpus.

> less than metaphorical sense. you know like he was using all these metaphors to describe, like sexual relations but, it's probably less likely that, it was like, him thinking about this is God's union with Israel. you know what i mean?

Nattinger and DeCarrico (1992) highlight this contrast in the presentation of academic registers and note that

> texts for second language learners, especially writing and reading texts, usually provide more opportunity for learning the discourse markers necessary for the comprehension of written discourse. What is lacking is this same opportunity for learning spoken discourse markers. Undoubtedly, this gap occurs because our second language research has thus far not provided us with enough information concerning these structures. (p. 80)

To learn more about the connections between the academic language that students read in assigned texts to prepare for their classes and the spoken version of academic English that students hear from their instructors and peers in class, we investigated the most commonly used four-word lexical bundles that appear in AWL and in MICASE and their different realizations in speaking and writing.

Lexical bundles are defined by Cortes (2004) as a "sequence of three or more words that co-occur frequently in a particular register" (p. 397) and include sets such as *I don't know why* or *in the case of.* Biber and Conrad (1999) propose that the difference between lexical bundles and other lexical groupings such as "lexical phrases," "formulas," or "prefabricated units" is that these are complete structural units or fixed expressions, whereas lexical bundles tend not to be complete structural units. This structural difference also distinguishes the investigation of lexical bundles from previous studies of lexical phrases in the academic register (Flowerdew & Tauroza, 1995; Nattinger & DeCarrico, 1992; Shaw, 1994). Before moving to the analysis of the corpus data, we begin with a discussion of the motivations behind our selection of particular corpora and some of the challenges that working with these corpora bring to researchers.

Working with AWL and MICASE

Selection and Use of AWL and MICASE

For this project we wanted to use two corpora of academic English: one of written academic English and the other of spoken academic English. Research using corpus linguistic methods is hampered by the lack of public access to many of the corpora that are used in published research. Some corpora such as the *Longman Spoken and Written English* corpus (used in Biber, Johansson, Leech, Conrad, & Finegan, 1999) or the TOEFL® 2000 Spoken and Written Academic Language (T2K-SWAL) Corpus (used in Biber, Conrad, Reppen, Byrd, & Helt, 2002) are privately owned or controlled with access limited to a few scholars working on projects sanctioned by the owner of the corpus. Other corpora such as the Bank of English or the British National Corpus are available for researchers who can afford to pay subscription fees for access or for copies provided on CDs or for copies downloaded from the Internet. Still other corpora are developed by researchers for their own purposes with access restricted either because copyright permissions limit who can use the materials or copyright permission has not been sought and the developer quite reasonably does not want to take chances with legal results from sharing a corpus with somewhat shaky legal standing. Fortunately, some institutions and government agencies have moved ahead with the development of high-quality corpora that are freely and easily available for all scholars who can access the Internet. Among these is the MICASE, an excellent resource for research into academic English not just because it has been made freely available by the University of Michigan on the Internet, but because it provides transcripts of spoken academic English in addition to some sound files. The MICASE corpus comprises 190 hours (approximately 1.7 million words) of academic speech from 152 speech events at the University of Michigan (Simpson-Vlach & Leicher, 2006). A broad sample of the kinds of speech that occur in academic settings is represented including classroom events (small and large lectures, lab and discussion sections, student presentations) and non-classroom events (advising sessions, office hours, dissertation defenses, colloquia). We wanted to use MICASE because of its free public access for other researchers who might want to follow up on this current study; however, we ourselves worked both with the MICASE website and with transcripts and sound files on CD-ROM that had been purchased by our department for faculty and student research.[2]

Working with the AWL Corpus

Selection of a corpus of written academic English was more challenging because we wanted data that would be widely accessible to other researchers. Eventually, we were able to get data from the corpus created by Coxhead

[2] We thank Dr. Gayle Nelson, Chair of the Department of Applied Linguistics & ESL at Georgia State University, for her support of our research through purchase of the MICASE database and the WM Recorder software.

to develop the AWL. Although access to that corpus is highly restricted, we wanted to use that source because the AWL has become widely used both in teaching materials and in research. Copyright restrictions meant that we could not see the entire AWL corpus, but we were able to get lexical bundle data and concordance lines of adequate but limited length. Similar access to the AWL will be made available to other researchers through the publication of a study of collocations and lexical bundles in the AWL (Coxhead, Bunting, Byrd, & Moran, forthcoming).

Working with MICASE

Written transcripts corresponding to all 152 speech events comprising the corpus are freely available at an online interface, *http://micase.umdl.umich.edu/m/micase*; however, the written transcriptions and sound files are accessed independently. Seventy sound files representing a portion of the written transcripts are freely available online at *www.lsa.umich.edu/eli/micase*. A further 61 sound files may be purchased from the University of Michigan on CD-ROM, and approximately 21 speech events are unavailable as sound files. This distribution of sound files between CD-ROMs and streaming audio files (or unavailable data) makes research focused on the speech data a somewhat onerous and time-consuming process. Once a search has been undertaken within the written transcription files and the needed transcript(s) identified, the researcher must then search through both the CD-ROM lists and online streaming files list to find out if the accompanying sound file is obtainable (as it may be one of the unavailable sound files) and if so, from where.

Additional challenges became apparent when we subjected specific clips of the sound files to instrumental measurement using a Kay Pentax Model 4300b Computerized Speech Laboratory (CSL) in order to determine the acoustic correlates (specifically pitch or fundamental frequency traces) of targeted words or bundles. The MICASE sound files were recorded using a digital audiotape recorder (DAT) with external microphones; hence, the quality of each recording varies depending on the amount of external noise in the room (chairs scraping, students chatting, equipment being moved, and so on). In order to clip portions of the online streaming files, we used WM Recorder, a software that captured the streaming files and converted them into .wav files. Sound clips from the CD recordings were exported in .wav format using Apple Quicktime. Sound files were then analyzed by the CSL for pitch. Where possible depending on sound quality, all data was subjected to both types of auditory and instrumental analysis.

Four-Word Lexical Bundles in AWL and MICASE

Work by Douglas Biber and his colleagues (Biber, Conrad, & Cortes, 2004) provides a solid description of the frequency of lexical bundles in written and spoken registers. In their comparisons of how these bundles operate in academic prose and conversations, bundles are classified structurally and functionally. Consistent differences are found between registers, most notably a tendency in conversation toward interactional or stance-related bundles such as *I don't know whether, I'll tell you what, you might as well,* as opposed to an inclination toward referential expressions with high transactional value in academic prose *(in the form of, in the case of, the way in which),* suggesting that "bundles have important discourse functions that fit the context and purposes of the registers in which they are common" (p. 13).

Biber, Conrad, & Cortes (2004) focus exclusively on academic discourse and find a similar division of lexical bundles by spoken and written register despite the higher informational load carried by spoken teaching discourse as opposed to conversation. They describe teaching discourse as "an intermediate register" containing both expressions of "personal concerns and interactions" and statements demonstrating the prioritizing of "informational goals" (p. 374). Thus, in our analysis we anticipated similar findings regarding the structure and function of lexical bundles in the AWL and MICASE corpora.

Although frequency counts of transcribed speech are the established measure employed to demonstrate the differences between modalities, it is the case that spoken discourse is understood by listeners as a series of acoustic signals, and this creates the potential for overlooking an essential characteristic of spoken corpora if the acoustic realizations of these lexical groupings are ignored. By analyzing only the written transcripts of spoken data, we fail to consider a crucial aspect of the differences between the presentation of language in spoken and written academic registers that L2 learners will have to contend with. It is well established that features of intonation and stress form a natural link between linguistic and sociolinguistic aspects of the language as they are paramount in both communicating a speaker's perception of the relative importance of the information load carried by different elements of the utterance in discourse contexts and in expressing relationships between discourse participants (Brazil, 1997; Crystal, 1969; Gumperz, 1982; Halliday, 1967; Tench, 1996; Williams, 1986). In inner-circle varieties of English (i.e., North American, British, Canadian, Australian, and New Zealand English), it is commonly understood that important items are distinguished from surrounding discourse by increased prominence, or pitch excursions and contour shapes, or tones, such as the common pedagogical distinction made between the rising tone at the end of yes-no questions (Are you ↗LEAVing?) and the falling tone at the end of wh- questions (Where are you ↘GOing?)

A typical example of how prominence is realized in discourse is shown, in which the contrasts between *acidity* and *alkalinity*, *lower* and *higher*, and *acid* and *base* are emphasized by clearly perceptible pitch peaks indicated in CAPS.

> We use the pH meter to MEAsure the aCIdity or ALkalinity of COMpounds/ if the pH value is LOwer than SEven then it's an ACid/ if the value is LARger than SEven then it's a BASE/ (Levis & Pickering, 2004)

Conversely, information that the speaker assumes to be within the hearer's current understanding may be "deaccented" (Wennerstrom, 1997, 2001)—that is, deliberately uttered low in the speaker's pitch range:

> '...for the BIcycle in the U.S. versus the $_{\text{BIcycle}}$ in CHIna...'
> In this utterance, the word *bicycle* is first introduced with a high accent, but in the second intonational phrase it has a low accent because it is already assumed to be in the hearers' consciousness (2001: 38)

Such cases of prosodic saliency within a specific discourse context are described by Brazil (1997). In his model of discourse intonation, Brazil proposes that the way in which syllables are assigned or selected for prominence rests on the pragmatic intentions of the speaker. The paradigm consists of what possible choices could appear in each of the syntagmatic slots of the utterance based on both the constraints of the language system (the general paradigm) and on the non-linguistic situation or the situated context of the interaction (the existential paradigm). For example, given a potential tone unit such as "a parcel of books lay on the table" at least two possible prominence selections could be made:

a. A parcel of BOOKS lay on the Table
b. A PARcel of books lay on the TAble

In (a) the speaker presents a prominent choice of "BOOKS" as opposed to perhaps flowers or cups and makes a similar prominence choice regarding the location—that is, on the table as opposed to on the floor or on the chair. The choice of prominence on both syllables projects a situated context in which both these pieces of information are unrecoverable either from prior interaction or from constraints within the language system. Equally, by choosing not to make prominent certain other words in the unit, the speaker assumes that no choice needs to be made for listener comprehension. This assumption may be based on linguistic or non-linguistic factors. For example, a choice of "box" of books (another possibility in the paradigm) can be considered synonymous with the choice of "parcel," and books can be assumed to "lay" on a table as opposed to "stand up." In terms of linguistic factors, typical constraints on possible choices in the language system apply to the nonprominent function words such as *of* and *on*, which are predetermined by the language system or content words that have already been negotiated (e.g., A: Was the book there? B: There was a PARcel of books there).

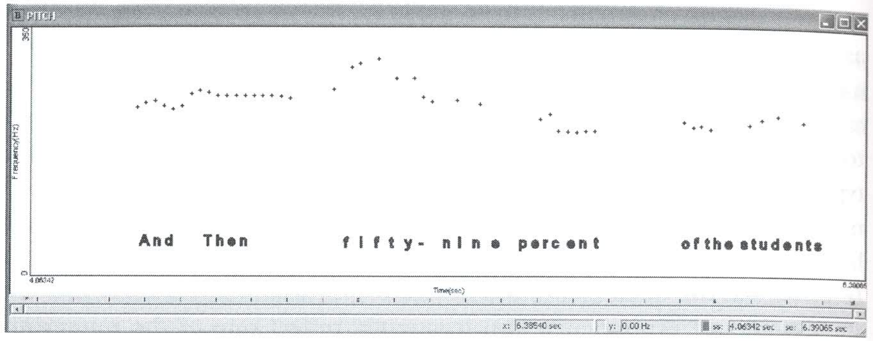
Choices of prominence also frequently reflect one of the numerous examples of conventionalized intonational idioms or routinized lexical phrases that currently operate in the language and that speakers believe their hearer(s) will be familiar with (Ashby & Ashby, 1994). An example of how a speaker's use of stress reflects his or her understanding of idiomatic patterns is demonstrated using prosodic realizations of percent taken from the MICASE corpus:

a. FIFty NINE per CENT
b. Can we ABsolutely a HUNdred percent say this is how it HAppened.

In (a) the speaker selects "perCENT" for prominence that distinguishes it from other possible choices such as "eighty cases" or "eighty times." Figure 6.1 shows the pitch peak [or fundamental frequency (FO)] exercursion associated with this syllable. In contrast, in (b) *percent* appears as part of an idiomatic phrase that the speaker treats as understood and realizes as non-salient, as shown in Figure 6.2. Similar prosodic realizations of *percent* occurred with "a HUNdred percent CHANCE of SUCcess" and "a HUNdred percent guaranTEE."

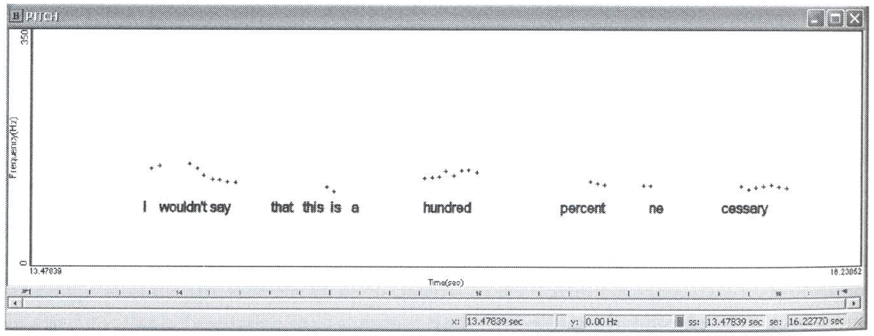
Non-linguistic factors affecting choice of prominence are multifarious and range from immediate discourse context (e.g., an item being currently within view of the hearer) to assumed cultural or situational knowledge. Brazil (1997) notes for example that in certain expressions such as "SAKS 5th AVEnue" and "BOTtle of johnnie WALker" alternatives to the nonprominent "5th" and "johnnie" "scarcely come to mind," and that these examples "demonstrate how conversationalists may rely on shared understandings that either have a very wide or very circumscribed currency" (pp. 24–25).

Figure 6.1
FO reading of a prominent realization of *percent* from the MICASE corpus



x: 6.38540 sec y: 0.00 Hz ss: 4.06342 sec se: 6.39065 sec

Figure 6.2
FO reading of a non-prominent realization of *percent* from the MICASE corpus



x: 13.47839 sec y: 0.00 Hz ss: 13.47839 sec se: 16.22770 sec

Additional examples of this kind of culturally assumed knowledge influencing choice of prominence include "SINgle white FEmale" and "LARGE metropolitan Area."

Within his discourse intonation model, Brazil also investigates the communicative function of tone or contour shape.[3] Using the principle of convergence, we can describe the social and informational significance of rising and falling contour shapes as the result of the speaker's assumptions regarding the current knowledge state of the hearer and the speaker's desire to project shared, common ground. Tone choice indicates the common ground between speakers in the following way: Falling tones project a speaker's assumption that the information is a new assertion and unrecoverable from the prior context, whereas rising tones project the assumption that the information is part of a shared background and is recoverable for the hearer from the prior discourse or assumed common knowledge. The following example, taken from Pickering (2001), demonstrates how this might be used in a teaching context. The teacher in this chemistry laboratory lecture chooses to "remind" students of the results of a flame test that they conducted a few weeks earlier using rising tones to emphasize the shared knowledge:

> //you remember that potassium was ↗PURple// . . . and the third possibility is that there be no color at ↗ALL//

Speakers make certain tone choices depending on the context they wish to project. The teacher could equally have selected falling tone choices and projected a context in which he was "telling" them again.

> //you remember that potassium was ↘PURple//...and the third possibility is that there be no color at ↘ALL//

The *deictic* function of the rising tone (Nevalainen, 1992) has high pragmatic significance in that it can be used to overtly signal solidarity (Pickering, 1999) and to avoid the appearance of direct contradiction (Hewings, 1995).

While we can presume that these choices are interpreted effortlessly by native speaking participants based on their shared sociocultural and linguistic experience, it is reasonable to assume that for non-native listeners without this background many of these choices will present a challenge. This is particularly the case as traditional EFL/ESL instruction focuses primarily on individual items and far less on the "phraseological character" of natural language (Cheng, 2006). Moreover, both teaching and research rarely focus on the spoken realization of any of these expressions.

Biber (2006) acknowledges the contribution of "paralinguistic marking" under which he subsumes pitch as well as gesture, body language, and facial expressions. However, he does not investigate it further in this study of university language. Prosodic characteristics and the choice of prominence or non-prominence on specific items are prioritized in Lindemann and Mauranen (2001). They focus on the word *just* as it functions in a portion of the MICASE corpus. In their consideration of the phonetic shape of the

[3] We only discuss the parts of Brazil's comprehensive discourse intonation model that are directly relevant to the analysis we report here. We do not believe that this distorts the foundational tenets of the model in any way; however, we refer the reader to the complete account of the model as it is presented in Brazil (1997).

word, including its prominence characteristics, the researchers found that when used as a "mitigator," *just* was generally realized as non-prominent, whereas when it operated as either an "emphasizer" or "particularizer," it was generally prominent, although there was some considerable overlap. Based on an informal study conducted by Lindemann and Mauranen (2001) to investigate the possibility of pragmatic misinterpretation by non-native listeners based on prominence patterns, they suggest that "in terms of teaching non-native speakers, this indicates a greater need to give attention to what might appear at first sight to be trivial matters (such as the phonetic form of *just*) in efforts to better prepare nonnative students for their socio-academic roles in USA universities" (p. 473).

Most recently, Warren (2006) specifically addresses prominence patterns in the Hong Kong Corpus of Spoken English, approximately half of which (one million words) has been prosodically transcribed. Also using Brazil's (1997) discourse intonation model, Warren traces lexical cohesive chains (Hoey, 1991) in job placement interviews (216 minutes of spoken data) and finds that prominence patterns are guided by what speakers perceive to be "situationally informative" (p. 311). In other words, prominence choices occur in response to local context and not as an intrinsic property of the word, a finding that supports Brazil's contention that prominence is speaker-selected and responsive to the immediate context of the interaction.

It seems clear then that non-native listeners will be at a considerable disadvantage when transferring information that they see on the page to comparable information that they hear. This obstacle may be even harder to overcome if, as Lindemann and Mauranen (2001) suggest, non-native listeners tend to focus on the lexical elements of the utterance and have difficulty integrating phonetic cues.

Procedure

Following Biber, Conrad, & Cortes (2004) and Cortes (2004), we limited our search to four-word lexical bundles and focused on the most frequently occurring lexical sequences of that length. The Wordlist tool from Wordsmith Tools 4.0 was used to develop lists of recurring four-word phrases: First, indexes were created with Wordlist for the complete MICASE and AWL corpora. These indexes were then used to search for four-word clusters[4] that were repeated

[4] The term *cluster* in Wordsmith Tools 4.0 corresponds to the *lexical bundle* used by Biber and his colleagues.

at least five times in the corpora. This approach provides data about lexical bundles that are characteristic of an entire corpus.[5] Various cut-off points have been used in the published literature on lexical bundles, based on the size of the corpora being analyzed. Generally, the larger a corpus, the more bundles that will be found, giving the researcher the ability to focus on highly frequent phrases at higher cut-off points. That is, there are no established standards for these cut-off points other than the general rule-of-thumb that the more frequent a phrase is, the more important it is for the corpus being studied. The usual practice is to norm the counts based on the appearance of clusters per million words. This norming makes it possible to compare the lexical bundles in corpora of different sizes like the AWL (3.8 million words) and the MICASE (1.7 million words). Because a cut-off of 10 per million words has been adopted in other published studies, we used that frequency for our basic standard and disregarded clusters that were repeated fewer times in the data.[6] Lexical bundles that involved proper nouns, such as *Victoria University of Wellington, the House of Lords,* or *the University of Michigan,* were left out of consideration, as were bundles that reflected special coding features of a corpus. Through this process of selection and elimination, we analyzed a total of 297 lexical bundles in AWL and 707 bundles in MICASE. This difference in numbers between the two registers accords with Biber, Conrad, & Cortes (2004) who found a much higher frequency of lexical bundles in classroom discourse as compared to written academic texts. Following previous work investigating lexical bundles in academic discourse, bundles in each corpus were classified by their structural and functional characteristics (Biber, 2006; Biber, Conrad, & Cortes, 2003; Biber, Conrad, & Cortes, 2004). In each case, because the number of cases of lexical bundles was so different between the two corpora, categories are shown as a percentage of the total.

Structural Characteristics

Studies of academic prose repeatedly report the prominence of long, complex noun phrases as the defining characteristic of such writing (e.g., Biber, 1988). Thus, the use of noun phrase fragments in the AWL as shown in Table 6.1 replicates previous studies as does the importance of prepositional phrase fragments since these are often post-nominal modifiers rather than

[5] Wordsmith Tools 4.0 can also create clusters from the Concord tool. These clusters focus on the behavior of particular words and phrases rather than on the linguistic characteristics of a corpus as a whole.

[6] Complete lists of clusters examined for the study are available from the authors.

Table 6.1
Structural Characteristics of Lexical Bundles in AWL and MICASE

	% Type 1 Verb Phrase Fragments	% Type 2 Dependent Clause Fragments	% Type 3 Noun Phrase & Prepositional Phrase Fragments
AWL	21	6	73
MICASE	54	20	26

free-standing adverbials. The contrast between the use of verb phrases in MICASE and in AWL accords with previous studies of these two types of academic communication: spoken academic communication involves both the verb-centric style of conversational English and the noun-centric style of academic writing in English. In addition to confirming the parallelism of these two corpora with other corpora and other reports of discourse analyses, this information reminds us strongly of the importance of selecting an appropriate corpus as the basis for discourse studies (see Baker, 2006, for a useful discussion of these issues).

Functional Characteristics

Overall results regarding the functional characteristics of lexical bundles are consistent with previous research into university language (Biber, 2006; Cortes, 2004). As shown in Table 6.2, in the AWL, referential bundles far outweigh any other functional type; in contrast, stance and interaction bundles, lexical groupings that express attitude or perform social functions in the discourse, account for a much larger proportion of lexical bundles in the MICASE corpus. In addition to the manifest difference in percentages in numbers of stance bundles, they were also qualitatively very different between the two modalities. Using the functional taxonomy developed in Biber, Conrad, & Cortes (2003, 2004) and Biber (2006), we find that in AWL, stance bundles appear as impersonal modality or attitudinal markers; typical examples are *is likely to be, it is possible to, should be able to, more likely to be, it may be that, would have to be.* By contrast, in MICASE the majority of these bundles appear as personal markers such as *I don't know what, oh I don't know, I think it was, you don't have to.* In this sense, the MICASE lexical bundles correspond more closely with lexical bundles found in conversation than in academic prose (Biber & Conrad, 1999). It is noteworthy that the functional category of text organizers comprises a similar percentage of the overall number of bundles found in AWL and MICASE. It is the case that university discourse has a high informational load regardless of modality and both teachers and textbook writers presumably strive to impose a clear organizational structure on that information using, in some part, text organizers.

Table 6.2
Functional Characteristics of Lexical Bundles in AWL and MICASE

	% Referential Bundles	% Stance & Interactional Bundles	% Text Organizers
AWL	72	17	11
MICASE	35	48	17

In order to examine the issues of the different functions performed by lexical bundles and their prosodic realizations in spoken discourse, we present sample analyses of four of the most frequent lexical bundles to appear in MICASE (compared with use of each phrase in AWL): (a) *at the same time, (b) at the end of*; (c) *I don't know I,* and (d) *you know what I.* In Biber, Conrad, & Cortes (2004), both *at the same time*and *at the end of*are classified as referential expressions. However, Biber (2006) recategorizes *at the same time* as a text organizer, and in this study, we look at the text organizing function of *at the end of.* These variations in analysis are one of the problematic aspects presented to the researcher when attempting to implement a system of functional categories. While Biber, Conrad, & Cortes (2004) acknowledge the possible problem of categorizing multifunctional bundles by addressing issues such as dual function and categorization by "most common use," this lack of exactness that is inherent in meaning-based systems remains an ongoing problem for data analysis. In some cases, as shown by Lindemann and Mauranen (2001), prosodic shape can be a crucial cue, and we further demonstrate this in this sample analysis of the lexical bundle *at the same time.*

1. *At the same time* (97.65 uses per million words in MICASE, 60.53 uses per million words in AWL)

Within MICASE, this bundle is only rarely used to indicate "real time" sequential action, for example:

> i mean, we hafta, we need to heat up, the the mash but, not **at the same time** that we'll be cooling down, uh, [S2**:** yeah cuz everything's like step by step] the mash

In many cases it functions rather as a text organizer highlighting the temporal simultaneity between two referential expressions, for example,

> Gregor Mendel, who we talked about, was writing, in the, later nineteenth century. he was ar- his work was around **at the same time** as Charles Darwin.

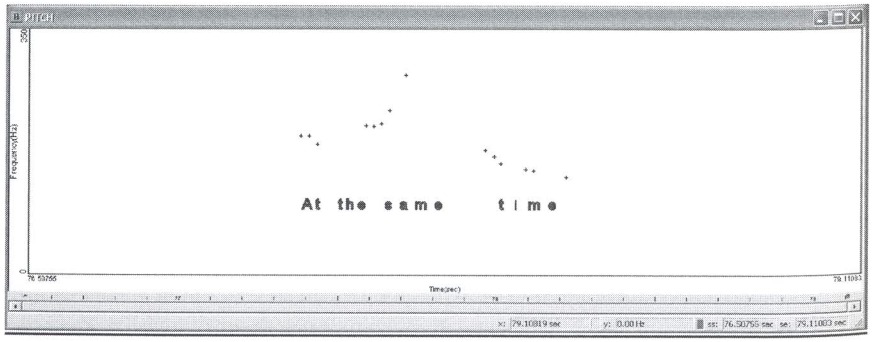
Most important, however, for any consideration of the functional purposes of the bundle, it is used to underscore a contrast between two aspects of a single idea or person:

> a. i guess that's the thing that makes Buddhism particularly Zen Buddhism, so uh fascinating and **at the same time** so frustrating, for a lot of people;
>
> b. this is something very important a- about Marx to keep in your mind. he_ his take on capitalism is not simple. yes capitalism sucks. and that's what most people know about Marx. right? that that's what he said. **at the same time,** capitalism is great. and these two words are_ sort of, encapsulate these ideas.

This key organizational function is reflected in the typically highly prominent pitch structure with which the lexical bundle is realized when it functions as a text organizer and that is shown in Fig 6.3. The prosodic shape comprises a pitch prominent peak "at the SAME time" and a robust falling tone; a crucial signal to the listener that this is an important assertion.

Figure 6.3

FO Reading of a Prominent, Falling (Assertive) Contour in the Realization of *at the same time* from the MICASE Corpus



2. *At the end of* (84.12 uses per million words in MICASE; 61.84 uses per million words in AWL)

At the end of is the sixth most frequent four-word lexical bundle in both the AWL and MICASE. Although categorized as a referential expression by Biber (2006), the phrase clearly fulfills a dual function in the MICASE corpus. In 60 percent of instances, it is used as a referential expression, for example:

> in the second century, **at the end of** the second century war had become a lot more complicated.

In the remaining 40 percent of cases, however, it is used as a discourse management marker and appears as part of classroom management talk—that is, as not indicating the organization of content but as indicating the organization of the class:

> a. you don't have the data and the outcome you can't write up your, discussion section, and so (i) should talk to (you) **at the end of** class.
>
> b. what i might wanna suggest is maybe **at the end of** the session i can show you a couple things. okay? mhm?
>
> c. alright we need a little calculator session **at the end of** of time today okay? cuz it is nice and easy to do it that way.

A useful distinction here is Coulthard and Montgomery's (1981) division of classroom content into main and subsidiary types. Main discourse comprises the informative content of the presentation, and the category of subsidiary content subsumes a variety of teaching discourse types concerned with class organization. Within spoken academic language, *at the end of* functions in both these capacities. The bundle demonstrates the difficulty of showing functional importance by use of frequency counts of a set phrase and suggests the need for a corresponding qualitative measure.

3. *I don't know I* (98.82 uses per million words in MICASE; 0 uses in AWL)

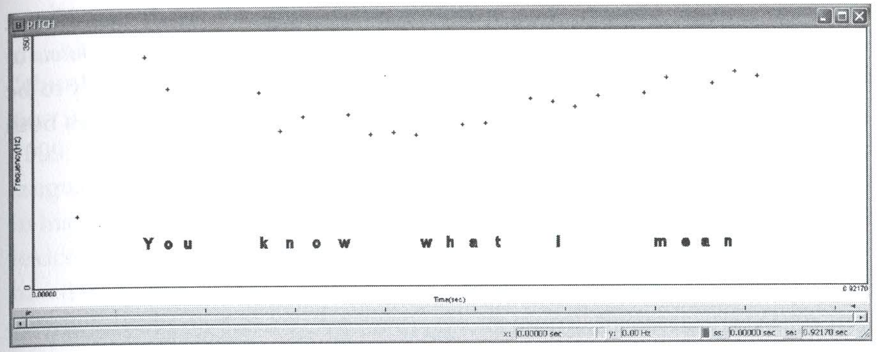
This lexical bundle frequently appears as a discourse device embedded in an utterance with other similar devices in strings such as *I don't know, I mean, I* *think it's* and is usually prosodically non-salient. This primary function as a "fluency device" (Nattinger & DeCarrico, 1992) is unique to the spoken modality, and its typical prosodic characteristics are a reflection of its importance for the speaker and corresponding unimportance in terms of semantic content for the hearer. This distinction, carried by both the lexical bundle and its prosodic realization, would seem to be key to an effective understanding of the discourse by a language learner. Thus, the verb *know* has become fixed in a phrase used by the instructor to hold the speaking space and to find time to think before speaking; that is, *know* is not any longer to be associated with epistemological stance but with a way of organizing spoken discourse in a manner that has become highly popular with large numbers of academics. In fact, we can demonstrate this difference using the Wordsmith Tools program to show that *know* appears 11,553 times (6795.88 per million) in the MICASE corpus, as opposed to 797 times (209.74 per million) in the AWL corpus. Thus, while we agree with Biber (2006) that a major purpose of university education is for teachers to help students learn to assess information and theory, we have found that the language of lexical bundles built around *know* might often need to be interpreted as text organizers or as teacher management bundles rather than as epistemic stance bundles.

4. *You know what I* (89.41 uses per million words in MICASE; 0 uses in AWL)

In contrast to *I don't know I*, this bundle appears most frequently in a prosodically salient pattern in which it collocates with *mean* or *I'm saying* to create the routinized expression *you know what I mean* or *you know what I'm saying*. This interactional device is considered to be an intonational idiom or "stereotype" (House, 1994) as it consistently appears with a final rising contour, for example, *you know what I* ↗ *MEAN*, *you know what I'm* ↗ *SAYing*, illustrated in Figure 6.4.

Figure 6.4

FO Reading of a Prominent, Rising (Solidarity-Building) Contour in the Realization of *do you know what I mean* from the MICASE Corpus



Particularly within classroom discourse, we recognize this kind of prosodic patterning as appearing on confirmation markers such as ↗ *Right* and ↗*OK* (Nattinger & DeCarrico, 1992; Pica & Long, 1986). Such devices, however, frequently do not function as genuine confirmation markers. When used for confirmation, the speaker pauses and waits for verification from listeners, while in classroom uses of such language the teacher does not pause and wait for verification from the students, turning the phrase into a rhetorical question. This excerpt from the MICASE corpus demonstrates how these phrases might appear in this kind of succession in an academic lecture:

> with Durer he had a little, knife, **you know what i mean?** he had a little knife and he was managing it with his hand right? and and and when he did etchings he was managing it again with his hand and his fingers, he had control over this part of his arm. now this is heavy labor. so um, to carve stone especially stone of this size, we're not talking about a Venus of Willendorf size stone which is, uh you can hold in your hand, here we're talking about, a block of marble that's bigger than the artist. right?

Rather than being "confirmation markers," they appear to operate as "solidarity markers" used by the teacher to implicitly acknowledge a negotiation between speaker and hearer and to create a sense of mutual participation in the discourse (Pickering, 2001). This use is both a fairly straightforward yet important characteristic of spoken discourse that a focus on the spoken realization of the corpus allows us to isolate.

In addressing the functional appearances of *know* as opposed to its grammatical categorization, we suggest that categories that prioritize grammatical realization are not sufficient to detect the significant, functional differences between different prosodic realizations of lexical bundles and thus the variety of pedagogical purposes they may perform. In addition, we suggest that functional categorization that does not recognize the multitude of different purposes that may prevail in one modality versus another—that is, spoken vs. written discourse—and that may be communicated outside the lexicogrammatical system—that is, by intonational structure—will obscure crucial communicative differences between spoken and written realizations of the same (or similar) phrases. This interpretation is suggested in the significant numerical difference between the frequency of the appearance of *know* in each corpus discussed here. Such significant differences are unlikely to be satisfactorily interpreted by one all-inclusive system that encompasses both spoken and written discourse.

Pedagogical Implications

It is evident that there are both connections and dissonances between what students will read in their writing assignments and hear in their classroom lectures and peer study groups. In agreement with Schleppegrell, Achugar, and Orteiza (2004), we believe students will need help to construct "a coherent message and knowledge framework from texts" (p. 74), and that teachers should assume that textual realization (i.e. as spoken or as written text) will require different kinds of help. In fact, we may need to illuminate the singular nature of classroom discourse, a complex and demanding register which comprises the most lexical bundles of each functional category and is "at the same time informational, involved and produced with real-time production constraints" (Biber, 2006, p. 148).

The AWL wordlists provide teachers and materials writers with access to individual words and word families that are characteristic of written academic discourse across the many disciplinary areas that make up a modern university (Coxhead, 2000). As Coxhead explains, these words are seldom more than 8 percent of an academic text, with much of the rest of the vocabulary in the text coming from the 1,000 most frequent words in English. Additionally, the most frequent of the most frequent words are structural words such as *the* and *of* rather than lexical forms such as nouns, adjectives, adverbs, or verbs. The implication is that the academic words are presented in phrases that often take a pattern such as *the .... of the ...* or *the ... of .....* Thus, learning to use the AWL words will mean learning to recognize, remember, access, and produce these longer strings. Teaching and learning vocabulary in phrases rather than just as individual words is a huge challenge because it goes against the traditional patterns of vocabulary instruction and demands information about how the words work in context. Another substantial challenge for teachers, materials writers, and curriculum designers is the yet unclear relationships and disconnections between the ways that academic words are

used in writing compared to their use in academic speech. Publication of information about the phrases and collocations found in the AWL corpus of written English will certainly be of help (Coxhead et al., forthcoming), but additional information will be needed about the use of these same words in spoken academic discourse.

Although a number of researchers have suggested that recognition of "chunks" or "formulaic sequences" may enhance second language acquisition in the classroom (Jones & Haywood, 2002; Myles, Mitchell, & Hooper, 1999; Sinclair, 1991), the potential saliency of lexical bundles for second language development is a relatively new area of discussion. Warren (2006) in his discussion of intonation units suggests that we look beyond traditional lexico-grammatical units in order to assess the significance of prosodically based units and argues that the study of discourse intonation should become an established part of English language teaching. He further suggests that corpora such as those considered here might be examined in order to "serve as the basis for learning and teaching materials" (p. 321).

It seems clear that students will require help in "transferring" written forms to their spoken counterparts, both in dealing with specialized vocabulary (Murphy, 2004) and in recognizing idiomatic units that may be ubiquitous in academic spoken language in relaying specific functions. A more recent publication that might form a template for this is Cauldwell's (2003) *Streaming Speech,* which works entirely with natural, spontaneous discourse and highlights the pitch, pause, and intonationally idiomatic characteristics of conversational English for learners both for listening and speaking purposes. In the same way, materials developers who are targeting EAP listening comprehension might choose authentic lectures (from corpora such as the MICASE) in which the kinds of expressions we have looked at here are abundant rather than cleaned up, error-free spoken prose that will not allow learners to develop the kinds of listening skills that will make them most effective in the classroom. Finally, in addition to providing students with work on understanding authentic academic speech, materials and lessons need to help students learn to make the connections between what they read and what they hear in class.

REFERENCES

Ashby, P., & Ashby, M. (1995). Spelling aloud: a preliminary study of idiomatic intonation. In J. Windsor-Lewis (Ed.), *Studies in general and English phonetics* (pp. 145–154). London: Routledge.

Baker, P. (2006). *Using corpora in discourse analysis.* London: Continuum.

Biber, D. (1988). *Variation across speech and writing.* Cambridge, UK: Cambridge University Press.

———. (2006). *University language.* Amsterdam: John Benjamins.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgard & S. Oksefjell (Eds.), *Out of corpora: Studies in honour of Stig Johansson* (pp. 181–190). Amsterdam: Rodopi.

Biber, D., Conrad, S., & Cortes, V. (2003). Lexical bundles in speech and writing: An initial taxonomy. In A. Wilson, P. Rayson, & T. McEnery (Eds.), *Corpus linguistics by the lune: A festschrift for Geoffrey Leech* (pp. 71–92). Frankfurt: Peter Lang.

———. (2004). If you look at ...: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly, 36*(1), 9–48.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* Harlow, England: Pearson Education Limited.

Brazil, D. (1997). *The communicative value of intonation in English.* Cambridge, UK: Cambridge University Press.

Cauldwell, R. (2003). *Streaming speech.* Birmingham, UK: Speechinaction.

Cheng, W. (2006). Describing the extended meanings of lexical cohesion in a corpus of SARS spoken discourse. *International Journal of Corpus Linguistics, 11*(3), 189–208.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*(4), 397–423.

Coulthard, M., & Montgomery, M. (1981). *Studies in discourse analysis.* London: Thomas Litho Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238.

Coxhead, A., Bunting, J., Byrd, P., & Moran, K. (forthcoming). *The Academic Word List: Collocations and recurrent phrases.* Ann Arbor: University of Michigan Press.

Crystal, D. (1969). *Prosodic systems and intonation in English.* Cambridge, UK: Cambridge University Press.

Flowerdew, J., & Tauroza, S. (1995). The effect of discourse markers on second language lecture comprehension. *Studies in Second Language Acquisition, 17*(4), 435–458.

Gumperz, J. (1982). *Discourse strategies.* Cambridge, UK: Cambridge University Press.

Halliday, M. (1967). *Intonation and grammar in British English.* The Hague: Mouton.

Hewings, M. (1995). Tone choice in the English intonation of non-native speakers. *International Review of Applied Linguistics, 3*(3), 251–265.

Hoey, M. (1991). *Patterns of lexis in text.* Oxford, UK: Oxford University Press.

House, J. (1994). Intonational stereotype: A reanalysis. In J. W. Lewis (Ed.), *Studies in general and English phonetics* (pp. 211–229). London: Routledge.

Jones, M., & Haywood, S. (2002). Facilitating the acquisition of formulaic sentences: An exploratory study in an EAP context. In R. Reppen, S. M. Fitzmaurice, & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 268–300). Philadelphia: John Benjamins.

Levis, J., & Pickering, L. (2004). Teaching intonation in discourse using speech visualization technology. *System, 32*(4), 505–524.

Lindemann, S., & Mauranen, A. (2001). "It's just real messy": The occurrence and function of just in a corpus of Academic speech. *English for Specific Purposes, 20*(Supplement 1), 459–475.

Murphy, J. (2004). Attending to word stress while learning new vocabulary. *English for Specific Purposes, 23*(1), 67–83.

Myles, F., Mitchell, R., & J. Hooper. (1999). Interrogative chunks in French L2. *Studies in Second Language Acquisition, 21*(1), 49–80.

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching.* Oxford, UK: Oxford University Press.

Nevalainen, T. (1992). Intonation and discourse type. *Text, 12*(3), 397–427.

Pica, T., & Long, M. (1986). The classroom linguistic and conversational performance of experienced and inexperienced teachers. In R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 85–98). Rowley, MA: Newbury House.

Pickering, L. (1999). *An analysis of prosodic systems in the classroom discourse of native speaker and nonnative speaker teaching assistants.* Unpublished dissertation, University of Florida, Gainesville.

———. (2001). The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly, 35*(2), 233–255.

Schleppegrell, M. J., Achugar, M., & Orteiza, T. (2004). The grammar of history: Enhancing content-based instruction through a functional focus on language. *TESOL Quarterly, 38*(1), 67–93.

Shaw, P. (1994). Discourse competence in a framework for ITA training. In C. M. C. Myers (Ed.), *Discourse and performance of international teaching assistants* (pp. 27–51). Arlington, VA: TESOL.

Simpson-Vlach, R. S., & Leicher, S. (2006). *The MICASE handbook: A resource for users of the Michigan Corpus of Academic Spoken English.* Ann Arbor: University of Michigan Press.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford, UK: Oxford University Press.

Tench, P. (1996). *The intonation system of English.* London: Cassell.

Warren, M. (2006). Because of the role of er front office um in hotel: Lexical cohesion and discourse intonation. *International Journal of Corpus Linguistics, 11*(3), 305–323.

Wennerstrom, A. (1997). *Discourse intonation and second language acquisition: Three genre-based studies.* University of Washington, Seattle.

———. (2001). *The music of everyday speech.* Oxford, UK: Oxford University Press.

Williams, B. (1986). An acoustic study of some features of Welsh prosody. In C. Johns-Lewis (Ed.), *Intonation in discourse* (pp. 35–51). Kent, UK: Croom Helm Ltd.