Text B7.2
R. Simpson

Rita Simpson (2004) 'Stylistic features of academic speech: the role of formulaic speech.' In U. Connor and T. Upton (eds), *Discourse in the professions*. Amsterdam: Benjamins.

Simpson employs some key corpus analysis techniques to identify high-frequency formulaic expressions in the spoken MICASE corpus (see Unit C7) and compares them with three comparison corpora. Following a careful identification of target patterns, she conducts frequency counts and comparisons then examines some typically academic formulas more closely in context to determine their functions in academic speech.

Task B7.2.1: Before you read

➤ Consider what formulaic expressions you would expect to find frequently in academic speech and whether they might differ from those in other spoken contexts.

Task B7.2.2: As you read

➤ Once again, note Simpson's use of both quantitative and qualitative methods to gain a better understanding of this feature and the different interpretations it allows her to make about spoken academic discourse.

R. Simpson

The corpora

This research is based on MICASE, a spoken language corpus of approximately 1.7 million words (200 hours) of contemporary university speech recorded at the University of Michigan between 1997 and 2001. Speakers represented in the corpus include faculty, staff and all levels of students, and include both native and non-native speakers. The data collection for the corpus involved recording entire speech events sampled across student levels and academic divisions including a variety of non-classroom academic speech events as well as the more traditional academic speech genres such as lectures, seminars, and class discussions.

The quantitative part of this study begins with a comparison of the frequencies of formulaic expressions across several different corpora of speech. Three corpora were chosen for comparison purposes: the Corpus of Spoken Professional American English (CSPAE), the Bank of English National Public Radio subcorpus (NPR), and the Switchboard Corpus (SWB). These corpora were chosen because they were the only sizable corpora of spontaneous spoken American English available at the time of the study. Of these three corpora, the one that is most similar to MICASE in terms of speech genre is CSPAE. This is a two-million-word corpus consisting of one million words of speech from White House press conferences and one million words of faculty committee meetings. The NPR corpus consists of over three million words of news radio broadcasts from National Public Radio. Switchboard is a corpus of casual phone conversations, approximately thirty minutes each, recorded between strangers who

were recruited specifically for the purpose of constructing the corpus and were given suggested topics of conversation. As it is the only corpus containing casual conversation, it is important for comparative purposes; but since it represents an unusual, contrived situation, it is less than ideal as an example of naturally occurring speech.

Analytical procedures

The methods of analysis used in this study are firmly grounded in a corpus-based approach. This approach involves, first of all, a text analysis program that can generate frequency statistics for sequences of words in the corpus, and secondly, a concordance program that shows all the occurrences of a particular phrase in its surrounding context. Using these methods allows for a detailed comparison of different genres based on quantitative evidence, and also permits more in-depth qualitative analysis of certain items chosen on the basis of those quantitative findings. Ultimately, the most revealing insights into professional discourse – or any particular language genre – will be gained from a closer look at the texts, the speakers, and the situational variables; quantitative analysis alone can never provide a satisfactory picture, especially when one of the goals of the research is to make the findings applicable to language teaching.

The units of analysis for this study were frequently occurring expressions of three, four or five words, which I refer to as high frequency formulate expressions. The minimum frequency used as a cutoff point was twenty tokens per million words (or thirty-four total tokens in MICASE).

In addition to this minimum frequency level as a basis for selecting which formulaic expressions to analyze, I applied the notions of structural and idiomatic coherence to further narrow the set of expressions investigated. Structural coherence refers to the syntactic composition of the word string; idiomatic coherence is essentially an intuitive notion. So only strings that constitute complete syntactic units, sentence stems, or that intuitively look, sound, and feel like idiomatically independent expressions were included in the set. Examples of syntactically complete units include prepositional phrases (*at the end*, *in the past*), noun phrases (*a lot of people*, *the first thing*, *something like that*), verb phrases (*to make sure*, *look at this*), or entire clauses (I *can't remember*, *does that make sense*). Examples of sentence stems include: I *think that*, I *don't have*, and *do you know*. And examples of idiomatically independent expressions include discourse marker strings such as *well you know*, or focusing expressions such as *the thing is*, or *it turns out* (*that*).

The entire list of three-, four-, and five-word strings in MICASE occurring at least twenty times per million words or a total of at least thirty-four tokens in the entire corpus included almost 1,800 expressions (1,611, 157, and eleven three-, four-, and five-word strings, respectively), but of these, only 224 expressions were classified as structurally coherent or idiomatically complete. . . . The final part of the analysis involved identifying, on the basis of the above comparisons, a few expressions that appeared to be quintessential academic formulae, and examining them in context from a pragmatic perspective.

Cross-corpus comparative frequencies

As already stated, the first step in this research was to find out which expressions occur most frequently in MICASE, and of these, which expressions are more frequent than

in the three comparison corpora. Table 1 shows the twenty most frequent three-, four-, and five-word expressions found in MICASE using the criteria discussed above. A number of these expressions, however, are also very frequent in other spoken corpora. So, in order to find out which expressions are typical of academic speech in particular, and not just characteristic high frequency expressions in any speech genre, I looked for the expressions that were significantly more frequent in MICASE than in all three of the comparison corpora, and these are listed in Table 2.

Table 1 Most frequent three, four and five-word formulaic expressions in MICASE

Expression	*Total tokens*	*Frequency (million)*	*Expression*	*Total tokens*	*Frequency (per million words)*
I don't know	1519	882	in other words	229	133
a little bit	669	389	at the end	229	133
in terms of	550	319	something like that	220	128
I don't think	503	292	and so on	216	125
I think that	482	280	do you know	212	123
you can see	368	214	what I mean	194	113
and I think	328	191	I don't have	179	103
do you think	258	150	the same time	173	101
I don't know if	256	149	but I think	173	101
the same thing	235	137	in this case	165	96

Table 2 Top twenty expressions significantly more frequent in MICASE than in all three comparison corpora

	Frequency (per million words)			
Expression	*MICASE*	*CSPAE*	*NPR*	*SWB*
you can see	214	41	26	36
and so on	125	59	28	23
what I mean	113	9	4	49
in this case	96	34	34	8
I was like	85	1	3	31
look at it	85	63	7	52
you don't know	84	31	11	43
so you have	82	19	7	35
point of view	77	55	28	24
you know what I mean	75	2	1	33
all of these	71	38	24	12
the first one	67	33	8	32
so we have	65	36	5	35
what I'm saying	64	23	4	27
look at this	60	32	8	6
and in fact	59	19	17	24
in the book	59	7	12	3
it doesn't matter	57	5	5	26
do you see	47	24	9	14

Analysis of selected expressions in context

In this section I turn to a small selection of phrases for a more detailed analysis of their contextual environments in order to further elucidate their functions in academic speech. These expressions were chosen on the basis of the results of the quantitative analysis as well as the range and salience of the functions they seem to be performing.

I'*m gonna* (*going to*) *go*

The first expression in this section is one that initially seemed unlikely to appear on a list of academic formulaic expressions; it is not obvious at first glance why I'*m gonna go* would be comparatively more frequent in academic speech. However, a look at the fifty examples from MICASE shows that nearly half of the uses of this expression have to do with discourse or task management, as in the expressions I'*m gonna go over/through/into* (*something*), meaning to discuss or present something in the class. The examples below illustrate this use:

(1) I'*m gonna go* through and give some examples.

(2) if I have time I'*m gonna go* over question three and five from the problem set.

Other similar uses have more to do with task management or the immediate sequencing of the unfolding discourse, as in these examples:

(3) I'*m gonna go* to roman numeral twenty-eight.

(4) I'*m gonna go* back and say something that I forgot to say.

the thing is

This expression in its discussive sense functions pragmatically as a focuser, prefacing and drawing attention to the ensuing comment or statement. However, a closer examination of the contexts in which the phrase occurs reveals a more complex pragmatic profile. First, it is often used when negating, contrasting, or qualifying – and simultaneously emphasizing – a crucial point:

(5) so I'm moving with the velocity here. but *the thing is* I'm not moving with the average velocity, right?

(6) *the thing is* here we are not doing the T-star version we're not going further and going through that cuz . . .

It is also used for explaining a problem, complication, or complex situation:

(7) *the thing is* that, that you have to make the sculpture so it can be free standing. that's a kind of a problem. you've gotta get it balanced right.

(8) *the thing is*, maximum size is, i- is rather a nebulous thing and it's rather difficult to determine.

Perhaps the most interesting usage is illustrated by the longer excerpt in (9), in which this expression is used while arguing a point interactively. Excerpt (9) is an example from a composition class of a student struggling with a small detail about rules of punctuation, in which she questions arid challenges the instructor. He in turn responds to her question 'A*re you sure*?' by launching into a slightly more detailed explanation, and prefaces the crux of the argument with *the thing is*, in order to draw attention not only to the content of the following point, but also to his conviction about the importance and validity of his explanation.

(9) Instructor: uhuh. this stuff goes inside, unless you've got a citation to include in your sentence [Student: okay.] this stuff goes outside.
Student: of quotations?
Instructor: right.
Student: always?
Instructor: always.
Student: see that's totally new to me. are you sure?
Instructor: [LAUGHS] it isn't actually. [Studs: LAUGH] um, here's why uh you can, {arrow} *the thing is* if you add a comma here and it's your comma and not Foucault's comma, you know you still need the comma so, it's alright. right? th- y- that's like it's sliding. it's it's – technically you're adding something to Foucault's text.

Discussion

The research for this study began by identifying a list of all three-, four-, and five-word formulate expressions in MICASE occurring above a specified frequency range. Following from that, approximately one-fourth of the expressions from that list were found to be significantly more frequent in MICASE than in three comparison corpora of other speech varieties, and thus particularly characteristic of academic speech. Finally, this research has examined the high frequency, characteristically academic formulaic expressions from a functional pragmatic perspective, showing that the most common functions can be broken down into two broad categories – functions related to the organization and structuring of discourse, and functions related to interactivity. There is a constant interplay between these two overarching characteristics of academic speech, which is by nature an information-rich genre, but in which interaction between the participants is also of paramount importance, and the formulaic expressions identified here serve to highlight these dual pragmatic features.

All of these expressions are valuable items for EAP students to learn both for listening as well as speaking. And, since they occur across the whole range of academic divisions, they need not be presented in subject-specific classes or contexts. These phrases are used as discourse structuring or organizing devices; for demonstrating, emphasizing, and hedging; for interactional purposes; and also sometimes as fillers. They are often crucial linking phrases between segments of the propositional content of utterances. As such, they contribute to idiomaticity and fluency in multiple ways, and are thus important items to include in an EAP curriculum.