An Academic Formulas List:
New Methods in Phraseology Research

Rita Simpson-Vlach and Nick C. Ellis

University of Michigan

rsimplach@gmail.com ncellis@umich.edu

To appear in Applied Linguistics, 2010
Accepted June 22, 2009

Prepublication draft

Keywords: English for academic purposes, academic formulas list, statistical measures, identification methods, pedagogical applications

Abstract

This research creates an empirically derived, pedagogically useful list of formulaic sequences for academic speech and writing, comparable to the Academic Word List (Coxhead 2000), called the Academic Formulas List (AFL). The AFL includes formulaic sequences identified as (1) frequent recurrent patterns in corpora of written and spoken language, which (2) occur significantly more often in academic than in non-academic discourse, and (3) inhabit a wide range of academic genres. It separately lists formulas that are common in academic spoken *and* academic written language, as well as those that are special to academic written language alone and academic spoken language alone. The AFL further prioritizes these formulas using an empirically derived measure of utility that is educationally and psychologically valid and operationalizable with corpus linguistic metrics. The formulas are classified according to their predominant pragmatic function for descriptive analysis and in order to marshal the AFL for inclusion in English for Academic Purposes instruction.

An Academic Formulas List (AFL)

The aim of this research is to create an empirically derived and pedagogically useful list of formulaic sequences for academic speech and writing, comparable to the Academic Word List (hereafter AWL, Coxhead 2000). It is motivated by current developments in language education, corpus linguistics, cognitive science, second language acquisition (SLA), and English for academic purposes (EAP). Research and practice in SLA demonstrates that academic study puts substantial demands upon students because the language necessary for proficiency in academic contexts is quite different from that required for basic interpersonal communicative skills. Recent research in corpus linguistics analyzing written and spoken academic discourse has established that highly frequent recurrent sequences of words, variously called lexical bundles, chunks, multiword expressions (*inter alia*) are not only salient but also functionally significant. Cognitive science demonstrates that knowledge of these formulas is crucial for fluent processing. And finally, current trends in SLA and EAP demand ecologically valid instruction that identifies and prioritizes the most important formulas in different genres.

The AFL includes formulaic sequences, identifiable as frequent recurrent patterns in written and spoken corpora that are significantly more common in academic discourse than in non-academic discourse and which occupy a range of academic genres. It separately lists formulas that occur frequently in both academic spoken and academic written language, as well as those that are more common in either written or spoken genres. A major novel development this research brings to the arena is a ranking of the formulas in these lists according to an empirically derived psychologically valid measure of utility, called 'formula teaching worth'

(FTW). Finally, the AFL presents a classification of these formulas by pragma-linguistic function, with the aim of facilitating their inclusion in EAP curricula.

Background

Functional, *Cognitive Linguistic* and *Usage-based theories* of language suggest that the basic units of language representation are *constructions* -- form-meaning mappings, conventionalized in the speech community, and entrenched as language knowledge in the learner's mind (Barlow and Kemmer 2000; Croft and Cruise 2004; Goldberg 2006; Langacker 1987; Robinson and Ellis 2008; Tomasello 1998, 2003). Constructions are associated with particular semantic, pragmatic, and discourse functions, and are acquired through engaging in meaningful communication. Constructions form a structured inventory of a speaker's knowledge of the conventions of their language, as independently represented units in a speaker's mind. Native-like selection and fluency relies on knowledge and automatized processing of these forms (Ellis 2009; Pawley and Syder 1983).

Corpus Linguistics confirms the recurrent nature of these formulas (Biber et al. 1998; Hunston and Francis 1996; McEnery and Wilson 1996). Large stretches of language are adequately described as collocational streams where patterns flow into each other. Sinclair (1991; 2004) summarizes this in his *'idiom principle:'* "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments." (1991, p. 110). Rather than being a minor feature, compared with grammar, Sinclair suggests that for normal texts, the first mode of analysis to be applied is the idiom principle, as most of text is interpretable by this principle.

Comparisons of written and spoken corpora demonstrate that more collocations are found in spoken language (Biber et al. 1999; Brazil 1995; Leech 2000). Speech is constructed in real time and this imposes greater working memory demands than writing, hence the greater the need to rely on formulas: it is easier to retrieve something from long-term memory than to construct it anew (Bresnan 1999; Kuiper 1996).

Many formulaic constructions are non-compositional or idiomatic, like 'once upon a time', or 'on the other hand', with little scope for substitution (*'twice upon a time', *'on the other foot') (Simpson and Mendis 2003). Even those that appear to be more openly constructed may nevertheless be preferred over alternatives (in speech, 'in other words' > 'to say it differently', 'in paraphrase', '*id est*') with the demands of native-like selection entailing that every utterance be chosen from a wide range of possible expressions, to be appropriate for that idea, for that speaker, for that genre, and for that time. Natives and experts in particular genres learn these sequential patterns through repeated usage (Ellis 1996, 2009; Pawley and Syder 1983; Wray 1999, 2002). Psycholinguistic analyses demonstrate that they process collocations and formulas with greater facility than 'equivalent' more open constructions (Bod et al. 2003; Bybee and Hopper 2001; Ellis 2002a, b; Ellis et al. 2009; Ellis et al., 2008; Jurafsky 2002; Schmitt 2004). For example, in speech production, "items that are used together, fuse together" (Bybee 2003, p. 112): words that are commonly uttered in sequence become pronounced as a whole that is shortened and assimilated ('give + me' -> 'gimme'; 'I + am + going + to' -> 'I'm gonna', etc.). The phenomenon is graded -- the degree of reduction is a function of the frequency of the target word and the conditional probability of the target given the surrounding words (Bybee and Hopper 2001; Jurafsky et al. 2001).

EAP research (e.g., Biber and Barbieri 2006; Flowerdew and Peacock 2001; Hyland 2004, 2008; Swales 1990) focuses on determining the functional patterns and constructions of different academic genres. These analyses have increasingly come to be based on corpora representative of different academic fields and registers, such as the Michigan Corpus of Academic Spoken English (Simpson et al. 2002), with qualitative investigation of patterns, at times supported by computer software for analysis of concordances and collocations. But these studies need to be buttressed with quantitative information too, as in the case of vocabulary where there have been longstanding attempts to identify the more frequent words specific to academic discourse and to determine their frequency profile, harking back, for example, to the University Word List (West 1953). The logic for instruction and testing is simple – the more frequent items have the highest utility and should therefore be taught and tested earlier (Nation 2001).

The most significant recent developments in this direction have been those of Coxhead (2000). Her development of the AWL has had a significant impact on EAP teaching and testing because it collects words that have high currency in academic discourse by applying specific criteria of frequency and range of distribution in a 3.5-million-word corpus of academic writing representing a broad spectrum of disciplines. Because academic study puts unique demands on language learners, the creation of the AWL as a teaching resource filled a substantial gap in language education by providing a corpus-based list of lexical items targeted specifically for academic purposes.

Can the same principles of academic vocabulary analysis be applied to other lexical units characterizing academic discourse? Can the theoretical research on formulaic language, reviewed above, which demonstrates that contiguous multiword phrases are important units of language,

be likewise transformed into practical pedagogical uses (Lewis 1993; Nattinger and DeCarrico 1992; Schmitt 2004; Wray 2000)? Is an AFL equally viable?

A crucial factor in achieving this goal lies in the principles for identifying and classifying such units. The *lexical bundle* approach of Biber and colleagues (Biber et al. 2004; Biber et al. 1998), based solely on frequency, has the advantage of being methodologically straightforward, but results in long lists of recurrent word sequences that collapse distinctions that intuition would deem relevant. For example, few would argue with the intuitive claim that sequences such as 'on the other hand' and 'at the same time' are more psycholinguistically salient than sequences such as 'to do with the', or 'I think it was', even though their frequency profiles may put them on equivalent lists. Selection criteria that allow for intuitive weeding of purely frequency-based lists, as used by Simpson (2004) in a study of formulaic expressions in academic speech, yield much shorter lists of expressions that may appeal to intuitive sensibilities, but they are methodologically tricky and open to claims of subjectivity.

In this paper we present a method for deriving a list of formulaic expressions that uses an innovative combination of quantitative and qualitative criteria, corpus statistics and linguistic analyses, psycholinguistic processing metrics and instructor insights. Long lists of highly frequent expressions are of minimal use to instructors who must make decisions about what content to draw students' attention to for maximum benefit within limited classroom time. The fact that a formula is above a certain frequency threshold and distributional range does not necessarily imply either psycholinguistic salience or pedagogical relevance; common sequences of common words, such as 'and of the,' are expected to occur frequently. Psycholinguistically salient sequences, on the other hand, like 'on the other hand', cohere much more than would be expected by chance; they are "glued together" and thus measures of association, rather than raw

frequency, are likely more relevant. Our primary aim in this research is to create a pedagogically useful list of formulaic sequences for academic speech and writing. A secondary aim, however, is to discuss the statistical measures beyond frequency counts available for ranking formulaic sequences extracted from a corpus. The departure point for our research was dissatisfaction with a strictly frequency-based rank ordering of multi-word phrases on the one hand, and a frequency plus intuition-based ordering on the other hand, coupled with a need for relatively contained, manageable sets of multi-word expressions for use in classroom applications and teaching materials development. We used frequency as a starting point, but our approach is substantially more robust than the previous corpus-based methods for classifying multi-word formulas in that it encompasses a statistical measure of cohesiveness – mutual information – that has heretofore not been used in related research, in conjunction with validation and prioritization studies designed to provide insights into which formulas are perceived to be the important ones for teaching.

Methods

The corpora

Target Corpora. The target corpora of academic discourse included 2.1 million words each of academic speech and academic writing. The academic speech corpus was comprised of MICASE (1.7 million words) (Simpson et al. 2002) plus BNC files of academic speech (431,000 words) (British National Corpus 2006). The academic writing corpus consisted of Hyland's (2004) research article corpus (1.2 million words), plus selected BNC files (931,000 words) sampled across academic disciplines using Lee's (2001) genre categories for the BNC.[1] The speech

corpus was broken down into five sub-corpora and the writing corpus into four sub-corpora by academic discipline, as shown in Table 1.

[TABLE 1 NEAR HERE]

Comparison Corpora. For comparative purposes, two additional corpora were used. For non-academic speech we used the Switchboard (2006) corpus (2.9 million words), and for non-academic writing we used the FLOB and Frown corpora (1.9 million words) which were gathered in 1991 to reflect British and American English over 15 genres and to parallel the original LOB and Brown collections (ICAME 2006). FLOB and Frown were favored over their predecessors because the age of the texts is closer to the target corpus texts. The Switchboard corpus was chosen because it contains unscripted casual telephone conversations, and thus lies near the opposite end of the style spectrum from academic speech.[2]

Formula identification and MI

The first decision was what length of formulas we would include in the data. It is well known that 2-word phrases (bi-grams) are highly frequent and include many phrases that are subsumed in 3- or 4-word phrases; so we excluded 2-word sequences, to keep the data set to a more manageable size. Although recurrent 5-word sequences are comparatively rare, we decided to include them for the sake of thoroughness, thus including strings of 3, 4, and 5 words into the data set. The next decision was what frequency level to use as a cutoff. Previous research uses cutoff ranges between 10 and 40 instances per million words. Since our research goals included using other statistical measures to cull and rank the formulas, we wanted a less restricted data set

to start with, and so opted for the lowest frequency range used in previous research, namely 10 per million (Biber et al, 1999).

We began by extracting all 3- 4- and 5-grams occurring at least 10 times per million from the two target and two comparison corpora, using the program *Collocate* (Barlow 2004). These four data sets naturally included a great deal of overlap, but also substantial numbers of phrases unique to each corpus. The next step then was to collapse the overlapping data and collect frequency counts for each phrase appearing in any one of those four corpora (at the threshold level of 10 per million) for all the other corpora, for comparison purposes. The total number of formulas in this list was approximately 14,000.

From this master list, we wanted to determine which formulas were more frequent in the academic corpora than in their non-academic counterparts, because our goal was to identify those formulas that are characteristic of academic discourse in particular, in contrast to high-frequency expressions occurring in any genre. This is an important step that warrants additional justification. Just as the AWL omitted words that were in the most frequent 2,000 words of English, we needed a way to sift out the most frequent formulas occurring in both academic and non-academic genres. To accomplish this, we used the log-likelihood statistic to compare the frequencies of the phrases across the academic and non-academic corpora. The log-likelihood ratio (LL) is useful for comparing the relative frequency of words or phrases across registers and determining whether the frequency of an item is statistically higher in one corpus or sub-corpus than another (Jurafsky and Martin 2000; Oakes 1998; Rayson and Garside 2000). Those expressions found to occur statistically more frequently in academic discourse, using the LL statistic with a significance level of p=.01, comprise the basis for the academic formulas list. We separately compared academic vs. non-academic speech, resulting in over 2000 items, and

academic vs. non-academic writing, resulting in just under 2000 items. The overlapping items from these two lists were identified as the core formulas that appear frequently in both academic speech and writing.

Once these lists were obtained, cutoff values for distributional range across the academic sub-divisions of the corpora had to be established. The sub-corpora for academic speech were (see Table 1): Humanities and Arts, Social Sciences, Biological and Health Sciences, Physical Sciences and Engineering, and Other/non-disciplinary. For academic writing, the sub-corpora were: Humanities and Arts, Social Sciences, Natural Sciences and Medicine, and Technology and Engineering. The cutoff values we used were as follows: Expressions occurring primarily in speech had to occur at the 10 tokens per million level or above in *four out of five* of the academic divisions, resulting in a Spoken AFL of 979 items; expressions occurring primarily in writing had to occur at least 10 times per million words in *three out of four* academic divisions, resulting in a Written AFL of 712 items; and expressions occurring in both speech and writing had to occur at a level of 10 per million in at least *six out of all nine* subcorpora, resulting in a Core AFL of 207 items.[3] These range thresholds ensure that the AFL formulas are found across the breadth of academic spoken or written language and are thus relevant to general EAP, rather than to particular disciplines. Furthermore, the range ensures that the formulas on the list are not attributable to the idiosyncrasies of particular speakers or speech events.

Another important statistic we calculated for each of the strings was the mutual information (MI) score. MI is a statistical measure commonly used in the field of information science designed to assess the degree to which the words in a phrase occur together more frequently than would be expected by chance (Manning and Schuetze 1999; Oakes 1998). A higher MI score means a stronger association between the words, while a lower score indicates

that their co-occurrence is more likely due to chance. MI is a scale, not a test of significance, so there is no minimum threshold value; the value of MI scores lies in the comparative information they provide. The question we then posed is: To what extent are these corpus metrics of frequency and MI useful for ranking the formulas on a list?

High frequency *n*-grams occur often. But this does not imply that they have clearly identifiable or distinctive functions or meanings; many of them occur simply by dint of the high frequency of their component words, often grammatical functors. In addition, relying solely on frequency means that some distinctively useful but lower frequency phrases whose component words are highly unlikely to occur together by chance will not make it to the top of the frequency-ordered list. So frequency alone is not a sufficient metric.

High MI *n*-grams are those with much greater coherence than is expected by chance, and this tends to correspond with distinctive function or meaning. But this measure tends, in contrast to frequency, to identify rare phrases comprised of rare constituent words, such as many subject-specific phrases. So nor is MI alone a perfect metric for extracting phrases that are highly noteworthy for teachers, since it privileges low-frequency items. Tables 2 and 3 present a simple re-ordering by frequency and MI of the top 10 and bottom 10 phrases of the approximately 2000 original Academic speech and original Academic writing items to illustrate these points.

[TABLES 2 AND 3 NEAR HERE]

For the speech data in Table 2, we see that frequency prioritizes such phrases as ‘and this is’ and ‘this is the’ which seem neither terribly functional nor pedagogically compelling, while it satisfactorily relegates to the bottom the phrases ‘cuz if you’, and ‘um and this’. Instructors might, however, be interested in other low frequency neighbors such as ‘we’re interested in’, and

'think about how'. MI, on the other hand, privileges functional formulas such as 'does that make sense' and 'you know what I mean', though 'blah blah blah' and 'the University of Michigan' are high on the list too. The low priority items by MI such as 'the um the' and 'okay and the' do indeed seem worthy of relegation. For the written data in Table 3, frequency highlights such strings as 'on the other hand' and 'it is possible' (we think appropriately), alongside 'it has been' and 'but it is' (we think inappropriately), and pushes 'by the use' and 'of the relevant' to the bottom (appropriately), alongside 'it is obvious that' and 'in the present study' (inappropriately). MI, in contrast, prioritizes such items as 'due to the fact that' and 'there are a number of' (appropriately; indeed all of the top ten seem reasonable), and it (appropriately) relegates generally non-functional phrases such as 'to be of', 'as to the', 'of each of' etc. These tables represent just a glimpse of what is revealed by the comparison of a given list of formulas ordered by these two measures. Our intuitive impressions of the prioritizations produced by these measures on their own, as illustrated here, thus led us to favor MI over frequency. Ideally, though, we wanted to combine the information provided by *both* metrics to better approximate our intuitions and those of instructors, and thus to rank the academic formulas for use in pedagogical applications.

Our efforts to achieve this synthesis were part of a large validation study which triangulated corpus linguistic measures, educator insights and psycholinguistic processing measures. A full description of these investigations is available in Ellis et al. (2008). Because these details are available elsewhere, and because the primary aim of the present paper is to present the AFL items and their functional categorizations, we simply summarize the relevant parts of the procedures here.

Determining a composite metric to index Formula Teaching Worth

We selected a subset of 108 of these academic formulas, 54 from the spoken and 54 from the written list. These were chosen by stratified random sampling to represent three levels on each of three factors: *n*-gram length (3, 4, 5), Frequency band (High, Medium, and Low; means 43.6, 15.0 and 10.9 per million respectively), and MI band (High, Medium, and Low; means 11.0, 6.7, and 3.3 respectively). There were two exemplars in each of these cells.

We then asked twenty experienced EAP instructors and language testers at the English Language Institute of the University of Michigan to rate these formulas, given in a random order of presentation, for one of three judgments using a scale of 1 (disagree) to 5 (agree):

A. whether or not they thought the phrase constituted 'a formulaic expression, or fixed phrase, or chunk'. There were six raters with an inter-rater $\alpha = 0.77$.

B. whether or not they thought the phrase has 'a cohesive meaning or function, as a phrase'. There were eight raters with an inter-rater $\alpha = 0.67$

C. whether or not they thought the phrase was 'worth teaching, as a bona fide phrase or expression'. There were six raters with an inter-rater $\alpha = 0.83$

Formulas which scored high on one of these measures tended to score high on another: r AB = 0.80, $p < .01$; r AC = 0.67, $p < .01$; r BC = 0.80, $p < .01$). The high alphas of the ratings on these dimensions and their high inter-correlation reassured us of the reliability and validity of these instructor insights. We then investigated which of Frequency or MI better predicted these instructor insights. Correlation analysis suggested that while both of these dimensions contributed to instructors valuing the formula, it was MI which more strongly influenced their prioritization: r Frequency/A = 0.22, $p < .05$; r Frequency/B = 0.25, $p < .05$; r Frequency/C = 0.26, $p < .01$; r MI/A = 0.43, $p < .01$; r MI/B = 0.51, $p < .01$; r MI/C = 0.54, $p < .01$. A multiple

regression analysis predicting instructor insights regarding whether an *n*-gram was worth teaching as a bona fide phrase or expression from the corpus metrics gave a standardized solution whereby teaching worth = β 0.56 MI + β 0.31 Frequency. That is to say, when instructors judge *n*-grams in terms of whether they are worth teaching, both frequency and MI factor into their judgments, but it is the MI of the string – the degree to which the words are bound together – that is the major determinant.

These beta coefficients, derived from the 108 formula subset for which we had obtained instructor ratings, could then be used over the population of academic formulas which they represented to estimate from the two corpus statistics available for all formulas—the combined measures of MI and frequency—a Formula Teaching Worth (FTW) score that is a prediction of how instructors would judge their teaching worth. This score, like the MI statistic, does not provide a threshold cut-off score, but enables a reliable and valid rank ordering of the formulas, which in turn provides instructors and materials developers with a basis for prioritizing formulaic expressions for instructional uses. The FTW score, with its use of both frequency rank and mutual information score is thus a methodologically innovative approach to the classification of academic formulas, as it allows for a prioritization based on statistical and psycholinguistic measures, which a purely frequency-based ordering does not.

Results: The AFL and Functional Categorization

In the appendix (see electronic supplement) we present the Academic Formulas List grouped into the three sublists—Core AFL in its entirety, and the first 200 formulas of the Spoken AFL, and Written AFL. Since all three lists are sorted by the two-factor FTW score, providing the top 200 formulas for the two longer AFL components effectively distills them into the most relevant formulas.

Scrutiny of the lists also shows substantial overlap among some of the entries. Thus, for example, in Appendix A the Core AFL listing includes the *n*-grams *from the point of view*, *the point of view of*, *point of view*, *point of view of*, *the point of view*, etc. Since this degree of redundancy is not especially useful, and moreover takes up extra space, in our functional categorizations we collapsed incidences like these together into their common schematic core--in this case, (*from*) (*the*) *point of view* (*of*). We retained the original formulas in the Appendix A tables, but only collapsed them in Table 4, the functional categorization. We acknowledge that in so doing we have sacrificed some detail as to the specific configurations and functions of component phrases; however, the differences in pragmatic function of these formula variations are generally minor and the detail lost can easily be retrieved by looking at the fuller lists in the appendix.

The final stage of the analysis involved grouping the formulas into categories according to their primary discourse-pragmatic functions. For purposes of expediency as well as the anticipated pedagogical applications, we again included only those formulas from the Core AFL list and the top 200 from the Written AFL and the Spoken AFL lists. These functional categories—determined after examining the phrases in context using a concordance program—are not meant to be taken as definitive and exclusive, since many of the formulas have multiple functions, but rather as indications of the most salient function the phrases fulfill in academic contexts. In the following section we present an overview of the functional analysis, providing examples to illustrate some of the more important functions in context.

Rationale and overview of the functional categories

The purpose of the following classification is primarily pedagogical. An ordered list of formulas sorted according to major discourse-pragmatic functions allows teachers to focus on

functional language areas which, ideally, will dovetail with functional categories already used in EAP curricula. The creation of a functional taxonomy for formulaic sequences is an inherently problematic endeavor, as Wray and Perkins (2000: 8) point out, arguing that typologies such as those offered by Nattinger and DeCarrico (1992), among others, suffer from a proliferation of types and sub-types. This proliferation of categories does indeed make it difficult to distill the data into a compact functional model applicable across corpora and domains of use. In spite of these difficulties, however, we maintain that for pedagogical purposes, a functional taxonomy, however multi-layered or imprecise because of overlapping functions and multi-functional phrases, is nevertheless crucial to enhancing the usefulness of the AFL for teachers. As for pedagogical applications, this functional categorization of the AFL is intended primarily as a resource for developing teaching materials based on further contextual research around the items rather than a resource for teaching itself. Due to space constraints, we cannot present specific teaching suggestions here, but do reiterate that the formula in context is what is pedagogically relevant. The functional categorization of the AFL is an important resource, but nevertheless only a starting point.

Previous researchers have in fact already paved the way in this area; in particular, we credit the work of Biber et al. (2004) in this aspect of our study. The current classification scheme is an adaptation of the functional taxonomy outlined in their article, but with some important extensions and modifications. As in their study, we grouped the formulas into three primary functional groups: referential expressions, stance expressions, and discourse organizers.

Several functional categories in our classification scheme, however, are not in the Biber et al. taxonomy, and these should be mentioned here. Within the referential expressions group, we have added one category – namely, that of contrast and comparison. This is a common

functional category in EAP curricula, and with over 20 formulas it represents an important functional group of the AFL. For the category of stance expressions, a number of formulas represent two essential categories not explicitly named by Biber et al.: These are hedges and boosters, and evaluation. In addition, we have collapsed two of their categories (desire, and intention/prediction) into one, called volition/intention, since the AFL formulas in the two categories did not seem distinct enough in their discourse functions to warrant splitting them. Finally, the discourse organizers group is substantially expanded and modified from the Biber et al. grouping, with three important additional subcategories: metadiscourse and textual reference, cause and effect expressions, and discourse markers. Our functional classification is thus considerably more extensive than Biber et al.'s; we suspect that this may be due primarily to the fact that there are close to 500 formulas in this portion of the AFL, compared to fewer than 150 phrases included in their list of the most common lexical bundles. Finally, we reiterate that even though some of the formulas are multifunctional, we have nevertheless tried to align all of them with their most probable or common function.

[TABLE 4 ABOUT HERE]

Description and examples of the functional categories

The following section outlines the pragmatic functional taxonomy. Numbers in brackets refer to the total number of formulas in that category from the combined Core AFL and top 200 each from the Written AFL and Spoken AFL.

A. **Referential Expressions**

The largest of the three major functional groupings, the referential expressions category encompasses five sub-categories: specification of attributes, identification and focus, contrast and comparison, deictics and locatives, and vagueness markers.

1) Specification of attributes

    a) Intangible framing attributes [66]

The largest pragmatic sub-category for all AFL phrases is the specification of attributes – intangible framing devices. The majority of these phrases appear on the Core AFL list, indicating that these are clearly important academic phrases across both spoken and written genres. This category includes phrases that frame both concrete entities (as in A.1) and abstract concepts or categories (as in A.2).

A.1) … based on the total volume passing through each cost center

A.2) so even with the notion of eminent domain and fair market value…

There are close to 70 formulas in this category, and roughly half are composed of the structure *'a/the N of'*, sometimes with a preceding preposition, as in *as a function of, on the basis of,* and *in the context of*. Most of these formulas frame an attribute of a following noun phrase, but some frame an entire clause (A.3), or function as a bridge between a preceding verb and a following clause (A.4).

A.3) But another clear example of the way in which domestic and foreign policy overlaps is of course in economic affairs.

A.4) human psychology has evolved in such a way, as to allow us to make those kind of judgments that would normally be reliable.

b) Tangible framing attributes [14]

The second subcategory of attribute specifiers is that of tangible framing attributes such as *the amount of, the size of, the value of,* which refer to physical or measurable attributes of the following noun.

A.5) this is uh, what she found in terms of the level of shade and yield of coffee…

c) Quantity specification [26]

The final subcategory of attribute specifiers is closely related to the category of tangible framing attributes, and includes primarily cataphoric expressions enumerating or specifying amounts of a following noun phrase, as in *a list of, there are three, little or no, all sorts of.* Some of the quantity specifiers, however – e.g., *both of these, of these two* – are anaphoric, referring to a prior noun phrase (e.g., A.7).

A.6) From an instrumental viewpoint, there are three explanations worth considering.

A.7) It is the combination of these two that results in higher profits to the EDLP store.

2) Identification and focus [53]

The second most common functional category, with 53 formulas, is the sub-category of identification and focus, which includes typical expository phrases such as *as an example, such*

as the, referred to as and *means that the*, and also a number of stripped-down sentence or clause stems with a copula, auxiliary verb or modal construction, such as *it is not, so this is, this would be*. It is not surprising that this functional category figures prominently in academic discourse, since exemplification and identification are basic pragmatic functions in both academic speech and writing. In fact these phrases often occur in clusters, as in example A.9.

A.8) So many religions, such as the religion of Ancient Egypt, for instance…

A.9) so this would be an example of peramorphosis.

3. Contrast and comparison [23]

Many of the contrast and comparison phrases included explicit markers of comparison such as *same*, *different, or similar*. As mentioned earlier, this category is not included in Biber et al., but constitutes an important language function for EAP teaching purposes.

A.10) that's probably a prefix code as opposed to a suffix code.

4. Deictics and locatives [12]

The deictic and locative expressions are a small but important functional category, referring to physical locations in the environment (e.g., *the real world)* or to temporal or spatial reference points in the discourse (e.g., *a and b, at this point)* These formulas obviously reflect the provenance of the corpus, so the *University of Michigan, Ann Arbor*, and *the United Kingdom* all appear on this list because of the inclusion of both MICASE and BNC texts.

5. Vagueness markers [4]

There are only four phrases included in the AFL that are classified as vagueness markers, making it the smallest functional category. Furthermore, three of these phrases are limited to the Spoken AFL; only the phrase *and so on* appears in the Core AFL. Nevertheless, the frequency rates and formula teaching worth scores (FTW) show that these phrases are important; making vague references with these particular extenders is a common discourse function in academic speech. Interestingly, Biber et al. (2004) also only list three phrases in this category (which they call imprecision bundles), yet claim that it is a major subcategory of referential bundles; perhaps this claim is also based on frequencies. Note that the three phrases they list in this category (*or something like that, and stuff like that, and things like that*), do not appear in the AFL, because although they may indeed be frequent in academic speech, they were not sufficiently *more* frequent in academic speech as compared to non-academic speech to make the cut for the AFL.

B. Stance Expressions

Stance formulas include six functional subcategories, two of which – hedges and evaluative formulas – are additions to the Biber et al. taxonomy.

1) Hedges [22]

This category includes a number of phrases that have multiple functions, but whose hedging function seems paramount (e.g., *there may be, to some extent, you might want to*). All of these formulas express some degree of qualification, mitigation, or tentativeness (Hyland 1998). Other examples of hedges show clearly the tendency of these formulas to co-occur with other hedge words or phrases, as in B.1, where the formula is preceded by 'I mean, uh, you know'.

B.1) but the, there are the examples of, and and the examples in the Renaissance I mean, uh, you know Copernicus is to some extent a figure of the Renaissance.

2) Epistemic stance [32]

Epistemic stance formulas have to do with knowledge claims or demonstrations, expressions of certainty or uncertainty, beliefs, thoughts, or reports of claims by others.

B.2) so we're just gonna be saying let's assume that the two variabilities in the two populations are the same…

3) Obligation and directive [23]

Obligation and directive formulas are generally verb phrases directing readers or listeners to do or not do something, or to recall or attend to some observation, fact, or conclusion.

B.3) Why? Tell me what your thought process is.

4) Ability and possibility [29]

The ability and possibility formulas frame or introduce some possible or actual action or proposition. In the spoken genres, these formulas are often interactive phrases with the second person pronoun, as in *you can see, you can look at* and *you're trying to.*

B.4) We aren't gonna be able to predict all behaviors because chance variables play a big role.

5) Evaluation [13]

The sub-category of evaluation is another addition to the Biber et al taxonomy. Biber et al. included only two of these phrases and listed them under the category of impersonal obligation/directive (i.e., *it is important to, it is necessary to*). The AFL, however, includes several phrases that are clearly evaluative, without necessarily being directive, such as *the importance of, is consistent with, it is obvious that, it doesn't matter.* Furthermore, even those that are also directive we maintain function primarily as evaluators. Interestingly, of the thirteen phrases in this category, most are on the Written AFL; only one appears on the Core AFL (*the importance of*), and one on the Spoken AFL (*it doesn't matter*).

B.5) Much macrosociological theory emphasizes the importance of societal variation.

6) Intention/volition [11]

Most of the phrases in this category occur in the spoken genres, and express either the speaker's intention to do something, or the speaker's questioning of the listener's intention.

B.6) So let me just take this off momentarily and put my other chart back on.

C. Discourse Organizing Expressions

Discourse organizers in the AFL fall into four main subcategories: metadiscourse, topic introduction, topic elaboration, and discourse markers. Each of these functions involves either signaling or referring to prior or upcoming discourse. With the exception of the cause-effect subcategory of topic elaboration, all the discourse organizing expressions are more frequent in

the spoken genres. This is consistent with Biber's (2006) finding that discourse markers are rare in written compared to spoken academic genres.

1) Metadiscourse and textual reference [31]

The sub-category of discourse organizers with the largest number of phrases is the metadiscourse and textual reference category. As mentioned earlier, this functional category was not included in the Biber et al. taxonomy; most of the phrases we classified in this category were grouped in their study with the topic introduction/focus category (2004: p. 386). With no phrases on the Core AFL, these phrases are clearly differentiated between the spoken and written lists, thus indicating that metadiscourse formulas tend to be genre-specific.

C.1) The seven studies are summarized in the next section.

C.2) Yeah I was gonna say something similar to that.

2) Topic introduction and focus [23]

This category overlaps functionally to a certain degree with the referring expressions identification and focus category. The main difference is that the global discourse organizing function of introducing a topic is primary here, with the phrase often framing an entire clause or upcoming segment of discourse, while the local referential function of identification is more salient for the other category.

C.3) so the first thing we wanted to do was take a look at and see if in fact this compound can kill cancer cells.

3) Topic elaboration

The topic elaboration subcategory includes two groups: non-causal topic elaboration, and cause and effect elaboration. Both categories function to signal further explication of a previously introduced topic.

a) Non-causal [15]

Non-causal topic elaboration includes any phrase that is used to mark elaboration without any explicit causal relationship implied. This includes phrases that summarize or rephrase, as in *it turns out that* and *what happens is,* as well as interactive formulas and questions such as *see what I'm saying*, and *any questions about.*

C.4) and let's just look at birth rate, and what happens is we have inverse, density dependence...

b) Cause and effect [22]

The cause and effect formulas signal a reason, effect, or causal relationship. Although these are grouped as a subset of the topic elaboration formulas, they are an important functional group in and of themselves in academic discourse and for EAP teaching.

C.5) at this point in order to get fired you have to do something really awful.

C.6) As a result, research on the imposition of the death penalty in the United States has a long and distinguished history.

4) Discourse markers [14]

The discourse markers category includes two sub-types. Connectives, such as *as well as, at the same time, in other words,* which connect and signal transitions between clauses or constituents. Interactive devices and formulas include *thank you very much, yes yes yes,* and *no no no*, which are phrases that stand alone and function as responses expressing agreement, disagreement, thanks, or surprise.

C.7) Material data as well as functional principles must be taken into account for the physical design.

Discussion and Conclusions

Our methods and results suggest that formulaic sequences can be statistically defined and extracted from corpora of academic usage in order to identify those that have both high currency and functional utility. Firstly, as in prior research with lexis (Nation 2001) and lexical bundles (Biber 2006; Biber et al. 2004), we used frequency of occurrence to identify constructions that appear above a baseline threshold frequency and which therefore have a reasonable currency in the language as a whole. Then, as in prior research defining academic lexis (Coxhead 2000), we identified those that appear more frequently in academic genres and registers and across a range of disciplines as being particular to EAP.

But currency alone does not ensure functional utility. However frequent in our coinage, nickels and dimes aren't worth as much as dollar bills. So too with formulas. When we assessed the educational and psycholinguistic validity of the items so selected, we found that they vary in worth as judged by experienced instructors, and in their processability by native speakers. In the

present paper we showed that experienced EAP and ESL instructors judge multiword sequences to be more formulaic, to have more clearly defined functions, and to be more worthy of instruction if they measure higher on the two statistical metrics of frequency and MI, with MI being the major determinant. In our companion paper (Ellis et al. 2008) we report experiments which showed how processing of these formulas varies in native speakers and in advanced second language learners of English.

Next, therefore, we used these findings to prioritize the formulas in our AFL for inclusion in English for Academic Purposes instruction using an empirically derived measure of utility that is both educationally valid and operationalizable with corpus linguistic metrics. Our FTW score weighs MI and frequency in the same way that EAP instructors did when judging a sample of these items for teaching worth. When we rank ordered the formulas according to this metric, the items which rose to the top did indeed appear to be more formulaic, coherent, and perceptually salient than those ordered by mere frequency or MI alone, thus providing intuitive confirmation of the value of the FTW score. We used this ordering to inform the selection and prioritization for inclusion in EAP instruction of the Core and the top 200 Written and Spoken AFL formulas. This inclusion of MI for prioritizing such multi-word formulas represents an important advance over previous research.

We then analyzed these formulas for discourse function to show that many of them fall into coherent discourse-pragmatic categories with enough face validity to encourage their integration into EAP instruction when discussing such functions as framing, identification and focus, contrast and comparison, evaluation, hedging, epistemic stance, discourse organization, and the like. Our AFL is categorized in this way in Table 4, with the functions further explained and exemplified in our Results section. It is our hope that this functional categorization, along

with the FTW rank-ordered lists, will facilitate the inclusion of AFL formulas into EAP curricula, and that further work on the pedagogical value of the AFL will take these results as a starting point.

We recognize that there are other possible ways of going about this task, each with particular advantages and disadvantages. Biber et al's groundbreaking work in defining lexical bundles on the basis of frequency alone has served as a contrast for us throughout this paper. It showed how corpus analysis could be used to identify interesting EAP constructions. But it also showed how frequency alone generates too many items of undifferentiated value. Biber et al. (2004) included only four-word bundles because the same frequency cut-off would generate far too many lexical bundles to deal with if three- and five-word bundles were included; yet, as we show here, many of the important (and high FTW) words on our AFL are actually tri-grams. So too, many of the phrases in their high-frequency lexical bundles list don't appear in the AFL because while they gathered all strings of frequency in university teaching and textbooks, we used comparison non-academic corpora and the LL statistic to pull out only those phrases that are particularly frequent in academic discourse.

Our conclusions also stand in contrast to those of Hyland (2008) who argues that there are not enough lexical bundles common to multiple disciplines to constitute a core academic phrasal lexicon, and therefore advocates a strictly discipline-specific pedagogical approach to lexical bundles. Although we would not deny that disciplinary variation is important and worthy of further analysis, by using the metrics we did, we were able to derive a common core of academic formulas that do transcend disciplinary boundaries. Several factors that explain our divergent claims warrant mentioning. First, Hyland also analyzed only four-word bundles, whereas a glance at the top 50 Core AFL phrases shows the majority to be three-word phrases

(e.g., *in terms of, in order to, in other words, whether or not, as a result)*. Second, he used a higher cutoff threshold, whereas we started with a lower cutoff frequency; since our FTW score incorporates another statistic (MI) to insure relevance, the lower frequency range allowed us to cast a wider net without also prioritizing numerous less relevant formulas. Our research thus finds quite a number of core formulas common to all academic disciplines.

In closing, we are left with important conclusions relating to the complementarity of corpus, theoretical, and applied linguistics. Whatever the extraction method, there are so many constructions that there is ever a need for prioritization and organization. The current research persuades us that we will never be able to do without linguistic insights both intuitive and academic. While some of these can be computationally approximated, as in the use of range of coverage of registers, and statistics such as MI and frequency in our FTW metric here, functional linguistic classification and the organization of constructions according to academic needs and purposes is essential in turning a list into something that might usefully inform curriculum or language testing materials.

References

Barlow, M. 2004. Collocate. Houston: Athelstan Publications.

Barlow, M. and S. Kemmer, (eds.). 2000. *Usage based models of language.* Stanford, CA: CSLI Publications.

Biber, D. 2006. *University language*. Amsterdam: John Benjamins.

Biber, D. and F. Barbieri. 2006. 'Lexical bundles in university spoken and written registers.' *English for Specific Purposes.*26: 263-286.

Biber, D., S. Conrad, and V. Cortes, 2004. '"If you look at …": Lexical bundles in university teaching and textbooks.' *Applied Linguistics* 25: 371-405.

Biber, D., S. Conrad, and R. Reppen, 1998. *Corpus linguistics: Investigating language structure and use.* New York: Cambridge University Press.

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

Bod, R., J. Hay, and S. Jannedy, (eds.). 2003. *Probabilistic linguistics*. Cambridge, MA: MIT Press.

Brazil, D. 1995. *A grammar of speech.* Oxford: Oxford University Press.

Bresnan, J. 1999. Linguistic theory at the turn of the century. Plenary presentation. 12th World Congress of Applied Linguistics, Tokyo, Japan.

British National Corpus. 2006. from http://www.natcorp.ox.ac.uk/.

Bybee, J. 2003. 'Sequentiality as the basis of constituent structure' in T. Givón and B. F. Malle (eds.): *The evolution of language out of pre-language*. Amsterdam: John Benjamins.

Bybee, J. and P. Hopper, (eds.). 2001. *Frequency and the emergence of linguistic structure*. Amsterdam: Benjamins.

Coxhead, A. 2000. 'A new Academic Word List.' *TESOL Quarterly* 34: 213-238.

Croft, W. and A. Cruise. 2004. *Cognitive linguistics*. Cambridge: Cambridge University Press.

Ellis, N. C. 1996. 'Sequencing in SLA: Phonological memory, chunking, and points of order.' *Studies in Second Language Acquisition* 18/1: 91-126.

Ellis, N. C. 2002a. 'Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition.' *Studies in Second Language Acquisition* 24/2: 143-188.

Ellis, N. C. 2002b. 'Reflections on frequency effects in language processing.' *Studies in Second Language Acquisition* 24/2: 297-339.

Ellis, N. C. 2009. 'Optimizing the input: Frequency and sampling in Usage-based and Form-focussed Learning' in M. H. Long and C. Doughty (eds.): *Handbook of Second and Foreign Language Teaching* (pp. 139-158). Oxford: Blackwell.

Ellis, N. C., E. Frey, and I. Jalkanen. 2009. 'The Psycholinguistic Reality of Collocation and Semantic Prosody (1): Lexical Access' in U. Römer and R. Schulze (eds.): *Exploring the Lexis-Grammar interface* (pp. 89-114).. Hanover: John Benjamins.

Ellis, N., R. Simpson-Vlach, and C. Maynard. 2008. 'Formulaic Language in Native and Second-Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL.' *TESOL Quarterly* 42(3): 375-96.

Flowerdew, J. and M. Peacock, (eds.). 2001. *Research perspectives on English for Academic Purposes*. Cambridge: Cambridge University Press.

Goldberg, A. E. 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Hunston, S. and G. Francis. 1996. *Pattern grammar: A corpus driven approach to the lexical grammar of English*. Amsterdam: Benjamins.

Hyland, K. 1998. *Hedging in Scientific Research Articles*. Amsterdam: John Benjamins.

Hyland, K. 2004. *Disciplinary discourses: Social interactions in academic writing*. Ann Arbor: University of Michigan Press.

Hyland, K. 2008. 'As can be seen: Lexical bundles and disciplinary variation.' *English for Specific Purposes* 27: 4-21.

ICAME. 2006. from http://icame.uib.no/.

Jurafsky, D. 2002. 'Probabilistic Modeling in Psycholinguistics: Linguistic Comprehension and Production' in R. Bod, J. Hay and S. Jannedy (eds.): *Probabilistic Linguistics*. Harvard, MA: MIT Press.

Jurafsky, D., A. Bell, M. Gregory and W. D. Raymond. 2001. 'Probabilistic relations between words: Evidence from reduction in lexical production.' in J. Bybee and P. Hopper (eds.): *Frequency and the emergence of linguistic structure*. Amsterdam: Benjamins.

Jurafsky, D. and J. H. Martin. 2000. *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics.* Englewood Cliffs, NJ: Prentice-Hall.

Kuiper, K. 1996. *Smooth talkers: The linguistic performance of auctioneers and sportscasters*. Mahwah, NJ: Erlbaum.

Langacker, R. W. 1987. *Foundations of cognitive grammar: Vol. 1. Theoretical prerequisites*. Stanford, CA: Stanford University Press.

Lee, D. 2001. 'Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle.' *Language Learning & Technology* 5/3: 37-72.

Leech, L. 2000. 'Grammars of spoken English: New outcomes of corpus-oriented research.' *Language Learning* 50: 675-724.

Lewis, M. 1993. *The lexical approach: The state of ELT and the way forward*. Hove, UK: Language Teaching Publications.

Manning, C. D. and H. Schuetze. 1999. *Foundations of statistical natural language processing*. Cambridge, MA: The MIT Press.

McEnery, T. and A. Wilson. 1996. *Corpus linguistics*. Edinburgh, UK: Edinburgh University Press.

Nation, P. 2001. *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nattinger, J. R. and J. DeCarrico. 1992. *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Oakes, M. 1998. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Pawley, A. and F. H. Syder. 1983. 'Two puzzles for linguistic theory: Nativelike selection and nativelike fluency' in J. C. Richards and R. W. Schmidt (eds.): *Language and communication*. London: Longman.

Rayson, P. and R. Garside. 2000. Comparing corpora using frequency profiling. Proceedings of the workshop on Comparing Corpora held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000), Hong Kong.

Robinson, P. and N. C. Ellis, (eds.). 2008. *A handbook of cognitive linguistics and second language acquisition*. London: Routledge.

Schmitt, N., (ed.). 2004. *Formulaic sequences*. Amsterdam: Benjamins.
Simpson, R. 2004. 'Stylistic features of academic speech: The role of formulaic expressions' in T. Upton and U. Connor (eds.): *Discourse in the professions: Perspectives from corpus linguistics*. Amsterdam: John Benjamins.
Simpson, R., S. Briggs, , J. Ovens and J. M. Swales. 2002. "The Michigan Corpus of Academic Spoken English." The Regents of the University of Michigan, Ann Arbor, MI.
Simpson, R. and D. Mendis. 2003. 'A corpus-based study of idioms in academic speech.' *TESOL Quarterly* 3: 419-441.
Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Sinclair, J. 2004. *Trust the text: Language, corpus and discourse*. London: Routledge.
Swales, J. M. 1990. *Genre analysis : English in academic and research settings*. Cambridge: Cambridge University Press.
Switchboard. 2006, Auust 5, 2006. '*A User's Manual*.' from http://www.ldc.upenn.edu/Catalog/docs/switchboard/.
Tomasello, M., (ed.). 1998. *The new psychology of language: Cognitive and functional approaches to language structure*. Mahwah, NJ: Erlbaum.
Tomasello, M. 2003. *Constructing a language*. Boston, MA: Harvard University Press.
West, M. 1953. *A General Service List of English Words*. London: Longman.
Wray, A. 1999. 'Formulaic sequences 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. and M. R. Perkins. 2000. 'The functions of formulaic language: An integrated model.' *Language and Communication* 20: 1-28.

Table 1

Word counts by discipline for the Academic sub-corpora

Academic Speech		Academic Writing	
Discipline	Word count	Discipline	Word count
Humanities and Arts	559,912	Humanities and Arts	360,520
Social Sciences	710,007	Social Sciences	893,925
Biological Sciences	357,884	Natural Sciences/Medicine	513,586
Physical Sciences	363,203	Technology and Engineering	349,838
Non-departmental/other	159,592		
Total	2,153,770	Total	2,117, 869

Table 2

The top 10 and bottom 10 phrases of the original Academic speech items prioritized by frequency and by MI

Top 10 by frequency	Top 10 by MI
this is the	blah blah blah
be able to	trying to figure out
and this is	do you want me to
you know what	for those of you who
you have a	we're gonna talk about
you can see	talk a little bit
look at the	does that make sense
you need to	thank you very much
so this is	the university of Michigan
you want to	you know what i mean

Bottom 10 by frequency	Bottom 10 by MI
if you haven't	okay and the
so what we're	is like the
as well but	so in the
cuz if you	and so the
right okay and	the um the
um and this	is what the
think about how	this in the
we're interested in	that it's the
will give you	is it the
we can we	of of of

Table 3

The top 10 and bottom 10 phrases of the original Academic writing items prioritized by frequency and by MI

Top 10 by frequency	Top 10 by MI
on the other	due to the fact that
in the first	it should be noted
the other hand	on the other hand the
on the other hand	it is not possible to
in the united	there are a number of
but it is	in such a way that
can be seen	a wide range of
it has been	take into account the
is likely to	on the other hand
it is possible	as can be seen

Bottom 10 by frequency	Bottom 10 by MI
is sufficient to	to the case
weight of the	of each of
of the relevant	with which the
by the use of	as in the
the assessment of	it is of
by the use	is that of
of the potential	to that of
it is obvious that	as to the
in the present study	to be of
is obvious that	that as the

Table 4

The AFL categorized by function.

The table includes all 207 formulas of the Core List, the top 200 items of the Written AFL and the top 200 items of the Spoken AFL lists

A. Referential Expressions

1. Specification of attributes

a) Intangible framing attributes

Core AFL (Written & Spoken)			
[a/the] form of (as) a function (of) based on [a/the] focus on the form of the (from) (the) point of view (of) in relation to in response to (in) the case (of) in the context (of) in the sense (that)	(in) such a (way) (in) terms of (the) in which the is based on (the) nature of the of the fact (on) the basis (of) the ability to the concept of the context of the definition of the development of	the distribution of the existence of (the) extent to which (the) fact that (the) the idea that the issue of the meaning of the nature of (the) the notion of the order of the presence of (a)	the problem of the process of the question of the role of the structure of the study of (the) way(s) in (which) the way that the work of the use of with respect to (the)
Primarily Spoken			
it in terms of	the idea of	the kind of	this kind of
Primarily Written			
an attempt to [are/was] based on by virtue of degree to which depend([ing/s]) on the	in accordance with (the) (in) such a way that in terms of a in the absence of	in the course of in the form of in this case the insight into the	on the basis of the on the part of to the fact that with regard to

b) Tangible framing attributes

Core AFL (Written & Spoken)			
(as) part of [a/the] the amount of the area of	the change in the frequency of the level of	(the) part(s) of the the rate of the sum of	(the) size of (the) (the) value of (the)
Written AFL			
an increase in the	high levels of	over a period of	

c) Quantity specification

Core AFL (Written & Spoken)			
a list of a series of a set of	[a/large/the] number of and the second	both of these each of [the/these] of [the/these] two	of the second the first is there are three
Primarily Spoken			
all sorts of			
Primarily Written			
a high degree a large number (of) (a) small number (of) (a) wide range (of)	little or no in a number of in both cases in most cases	in some cases (the) total number (of) (there) are a number (of)	there are no there are several two types of

2. Identification & focus

Core AFL (Written & Spoken)			
a variety of [an/the] example of (a) as an example different types of here is that if this is	is for the is not [a/the] is that [it/the/there] is the case is to be it can be it does not	it is not means that the referred to as such as the that in [a/the] that is the that there [are/is (a)]	that this is that we are there is [a/an/no] this is [a/an/not] this type of this would be which is [not/the]
Primarily Spoken			
[has/have] to do with it's gonna be and this is for those of you (who)	how many of you nothing to do with one of these	so this is the best way to there was a	this is the this is this is those of you who
Primarily Written			
(as) can be seen (in) does not have has also been his or her	it has been none of these that it is not	that there is no there has been they [did/do] not	this does not this means that which can be

3. Contrast & comparison

Core AFL (Written & Spoken)			
and the same as opposed to associated with the between the two	different from the exactly the same have the same [in/of/with] the same	is much more related to the the same as	(the) difference between (the) the relationship between
Primarily Spoken			
(nothing) to do with (the)	the same thing	to each other	

Primarily Written			
be related to the is more likely	(on) the other (hand) (the) similar to those	the difference between the	(the) same way as to distinguish between

4. Deictics & locatives

Core AFL (Written & Spoken)			
a and b	the real world	of the system	
Primarily Spoken			
(at) the end (of) (the) at this point	(at) (the) university of Michigan	in ann arbor	piece of paper
Primarily Written			
at the time of	at this stage	b and c	the united kingdom

5. Vagueness markers

Core AFL (Written & Spoken)			
and so on			
Primarily Spoken			
and so forth	and so on and so	blah blah blah	

B. Stance Expressions

1. Hedges

Core AFL (Written & Spoken)			
(more) likely to (be)	[it/there] may be	may not be	to some extent
Primarily Spoken			
a kind of a little bit about in a sense	it could be it looks like	it might be little bit about	might be able (to) you might want to
Primarily Written			
appear(s) to be are likely to as a whole	at least in does not appear	is likely to (be) it appears that	it is likely that less likely to

2. Epistemic stance

Core AFL (Written & Spoken)			
according to the be the case	assume that the out that the	to show that	we can see
Primarily Spoken			
[and/as] you can (see) do you know what (does) that make sense	how do we how do you know I think this is	trying to figure (out) to figure out (what) you think about it okay I don't know	what do you mean what does that mean (you) know what I (mean)
Primarily Written			
assumed to be be argued that be explained by be regarded as	be seen as been shown to can be considered	be considered as have shown that if they are	is determined by we assume that we have seen

3. Obligation & directive

Primarily Spoken			
do you want (me) (to) doesn't have to be don't worry about has to be	I want you to it has to be keep in mind take a look (at)	tell me what (to) make sure (that) we have to we need to	you don't need to you need to (do) you want me to you want to
Primarily Written			
(it should) be noted (that)	need not be needs to be	should also be should not be	take into account (the) to ensure that (the)

4. Expressions of ability & possibility

Core AFL (Written & Spoken)			
can be used (to)	to use the		
Primarily Spoken			
(gonna) be able (to) so you can (see)	that you can to think about	(you) can look at you can see ([that/the])	you could you could you're trying to
Primarily Written			
allows us to are able to be achieved by [be/been/was] carried out carried out [by/in]	be used as a be used to can also be can be achieved can be expressed	can easily be can be found (in) could be used has been used (it) is not possible (to)	it is possible ([that/to]) most likely to their ability to to carry out

5. Evaluation

Core AFL (Written & Spoken)			
the importance of			
Primarily Spoken			
it doesn't matter			
Primarily Written			
important role in is consistent with it is difficult	it is important (to) it is impossible to it is interesting to	it is necessary (to) it is obvious that it is worth	(it) is clear (that) the most important

6. Intention/volition, prediction

Primarily Spoken			
I just wanted to I wanted to	if you wanna if you want(ed) (to)	if you were (to) I'm gonna go	I'm not gonna let me just um let me
Primarily Written			
to do so	we do not		

C. Discourse Organizing Functions

1. Metadiscourse & textual reference

Primarily Spoken			
come back to go back to the gonna talk about I was gonna say (I) was talking about I'll talk about	I'm talking about talk a little bit talk(ing) about the to talk about wanna talk about	we talk(ed) about we were talking (about) we'll talk about we're gonna talk (about) we're talking about	we've talked about what I'm saying what I'm talking about what you're saying you're talking about
Primarily Written			
as shown in at the outset in table 1	in the next section in the present study in this article	(in) this paper (we) shown in figure	shown in table the next section

2. Topic introduction & focus

Core AFL (Written & Spoken)			
for example [if/in/the]	what are the		

Primarily Spoken			
a look at first of all I have a question I'll show you if you have (a) if you look (at) (the)	if you've got let's look at look at [it/the/this] looking at the to look at (the)	wanna look at we look(ed) at we're looking at what I mean what I want to	when you look at you have a you look at (the) you're looking at you've got a

3. Topic elaboration

a) non-causal

Core AFL (Written & Spoken)			
but this is			
Primarily Spoken			
any questions about came up with come up with (a)	I mean if (you) (it) turns out (that)	see what I'm saying so if you	what happens is you know what I'm
Primarily Written			
are as follows factors such as	in more detail	see for example	such as those

b) Topic elaboration: cause & effect

Core AFL (Written & Spoken)			
[a/the] result of (as) a result (of) because it is	due to the in order to	so that the the effect(s) of	the reason for whether or not (the)
Primarily Spoken			
end up with	in order to get	the reason why	
Primarily Written			
as a consequence as a result of the due to the fact (that)	for the purposes of for this purpose for this reason	give rise to is affected by	it follows that to determine whether

4. Discourse markers

Core AFL (Written & Spoken)			
and in the	as well as	at the same (time)	(in) other words (the)
Primarily Spoken			
and if you and then you	but if you by the way	no no no (no) thank you very (much)	oh my god yes yes yes
Primarily Written			
even though the	in conjunction with		

Appendices

Appendix A: Core AFL Academic Formulas (sorted by 2-factor FTW score)

		Speech		Writing		
		Raw freq	Freq per million	Raw freq	Freq per million	FTW
1	in terms of	726	337	597	282	3.53
2	at the same time	198	92	208	98	2.56
3	from the point of view	23	11	30	14	2.44
4	in order to	330	153	540	255	2.35
5	as well as	147	68	539	255	2.08
6	part of the	432	201	457	216	1.96
7	the fact that	383	178	430	203	1.96
8	in other words	315	146	188	89	1.90
9	the point of view of	22	10	31	15	1.89
10	there is a	307	143	472	223	1.72
11	as a result of	57	26	158	75	1.58
12	this is a	657	305	160	76	1.57
13	on the basis of	50	23	174	82	1.50
14	a number of	190	88	455	215	1.50
15	there is no	107	50	391	185	1.45
16	point of view	177	82	128	60	1.41
17	the number of	171	79	521	246	1.38
18	the extent to which	21	10	111	52	1.36
19	as a result	131	61	264	125	1.35
20	in the case of	68	32	286	135	1.32
21	whether or not	128	59	172	81	1.31
22	the same time	207	96	221	104	1.26
23	with respect to	94	44	220	104	1.26
24	point of view of	27	13	32	15	1.22
25	as a function of	40	19	77	36	1.19
26	at the same	235	109	248	117	1.19
27	the point of view	36	17	32	15	1.13
28	in such a way	28	13	41	19	1.11
29	the use of	58	27	572	270	1.11
30	in other words the	47	22	39	18	1.08
31	in terms of the	120	56	142	67	1.07
32	more likely to	91	42	161	76	1.06
33	likely to be	79	37	243	115	1.03

34	in this case	188	87	193	91	1.03
35	as opposed to	115	53	84	40	1.02
36	the way in which	59	27	82	39	0.94
37	based on the	65	30	283	134	0.91
38	can be used	27	13	189	89	0.87
39	the relationship between	45	21	159	75	0.87
40	it is not	71	33	398	188	0.81
41	and so on	329	153	145	68	0.79
42	on the basis	52	24	199	94	0.75
43	the difference between	95	44	80	38	0.74
44	it may be	96	45	184	87	0.72
45	the presence of	46	21	276	130	0.70
46	in the sense that	73	34	47	22	0.70
47	a variety of	55	26	193	91	0.69
48	different types of	51	24	59	28	0.69
49	extent to which	21	10	114	54	0.66
50	exactly the same	92	43	35	17	0.65
51	a series of	50	23	184	87	0.63
52	in relation to	25	12	192	91	0.63
53	it can be	119	55	206	97	0.63
54	the case of	84	39	356	168	0.62
55	in the case	78	36	323	153	0.62
56	large number of	21	10	75	35	0.62
57	that there is a	49	23	87	41	0.61
58	to some extent	56	26	39	18	0.60
59	that there is	148	69	246	116	0.59
60	the real world	41	19	38	18	0.57
61	is based on	37	17	125	59	0.56
62	due to the	45	21	269	127	0.55
63	ways in which	49	23	79	37	0.54
64	an example of	132	61	86	41	0.54
65	the fact that the	32	15	122	58	0.54
66	referred to as	24	11	66	31	0.52
67	may not be	71	33	104	49	0.52
68	way in which	96	45	114	54	0.51
69	it does not	28	13	145	68	0.48
70	from the point of	23	11	31	15	0.47
71	the development of	57	26	256	121	0.46
72	in the same	192	89	187	88	0.46
73	a result of	70	33	182	86	0.46
74	the basis of	58	27	229	108	0.45
75	the role of	53	25	257	121	0.43
76	there may be	57	26	67	32	0.43
77	difference between the	33	15	70	33	0.42
78	between the two	71	33	143	68	0.41
79	the size of the	37	17	67	32	0.41

80	the importance of	43	20	199	94	0.40
81	that there are	136	63	134	63	0.39
82	as a function	41	19	77	36	0.34
83	associated with the	25	12	115	54	0.31
84	the amount of	107	50	140	66	0.30
85	a function of	74	34	128	60	0.29
86	as an example	24	11	45	21	0.27
87	for example if	42	20	41	19	0.26
88	such as the	25	12	223	105	0.26
89	based on a	35	16	76	36	0.26
90	as part of	46	21	130	61	0.25
91	this is not	145	67	132	62	0.25
92	in which the	58	27	352	166	0.24
93	the effect of	44	20	232	110	0.24
94	in response to	30	14	121	57	0.22
95	related to the	46	21	163	77	0.22
96	each of these	51	24	70	33	0.21
97	the effects of	39	18	211	100	0.21
98	terms of the	125	58	168	79	0.20
99	we can see	57	26	25	12	0.20
100	there are three	22	10	31	15	0.20
101	for example the	37	17	177	84	0.18
102	according to the	42	20	149	70	0.18
103	the existence of	31	14	133	63	0.18
104	the concept of	46	21	149	70	0.18
105	in this way	37	17	141	67	0.17
106	focus on the	36	17	60	28	0.16
107	the nature of	41	19	179	85	0.15
108	the context of	32	15	189	89	0.15
109	a list of	67	31	68	32	0.15
110	this type of	39	18	84	40	0.14
111	such a way	28	13	41	19	0.13
112	the ability to	53	25	114	54	0.13
113	the idea that	126	59	80	38	0.13
114	a set of	45	21	145	68	0.11
115	other words the	47	22	39	18	0.11
116	parts of the	69	32	115	54	0.09
117	nature of the	38	18	164	77	0.09
118	the level of	67	31	163	77	0.05
119	this would be	89	41	39	18	0.05
120	is that the	252	117	200	94	0.04
121	is much more	23	11	37	17	0.04
122	the same as	95	44	85	40	0.04
123	to show that	48	22	79	37	0.04
124	there is an	44	20	87	41	0.03
125	the notion of	39	18	104	49	0.03

126	in the sense	112	52	81	38	0.00
127	in the context	23	11	129	61	0.00
128	the process of	80	37	145	68	-0.01
129	is not a	93	43	171	81	-0.02
130	both of these	49	23	25	12	-0.03
131	for example in	24	11	77	36	-0.03
132	the part of the	28	13	50	24	-0.05
133	the size of	61	28	98	46	-0.06
134	the form of	50	23	165	78	-0.06
135	the sum of	39	18	66	31	-0.08
136	the reason for	35	16	62	29	-0.09
137	a and b	35	16	101	48	-0.10
138	that this is	196	91	69	33	-0.11
139	fact that the	36	17	133	63	-0.11
140	this is an	117	54	40	19	-0.12
141	because it is	48	22	86	41	-0.13
142	have the same	81	38	48	23	-0.15
143	part of a	42	20	103	49	-0.18
144	the question of	52	24	136	64	-0.19
145	of these two	38	18	43	20	-0.21
146	the value of	35	16	126	60	-0.21
147	assume that the	23	11	49	23	-0.21
148	size of the	40	19	83	39	-0.21
149	in such a	38	18	97	46	-0.21
150	the distribution of	22	10	82	39	-0.22
151	of the same	69	32	149	70	-0.22
152	the meaning of	21	10	84	40	-0.23
153	view of the	43	20	108	51	-0.23
154	each of the	38	18	162	77	-0.24
155	which is not	45	21	57	27	-0.25
156	the issue of	35	16	76	36	-0.25
157	but this is	93	43	51	24	-0.26
158	if this is	87	40	44	21	-0.27
159	the rate of	30	14	100	47	-0.27
160	that we are	64	30	63	30	-0.31
161	with the same	44	20	52	25	-0.31
162	the result of	23	11	100	47	-0.31
163	the problem of	41	19	117	55	-0.31
164	is to be	35	16	181	85	-0.32
165	the study of	25	12	124	59	-0.32
166	which is the	153	71	72	34	-0.33
167	the definition of	34	16	49	23	-0.33
168	here is that	81	38	23	11	-0.35
169	from the point	24	11	34	16	-0.35
170	a form of	25	12	63	30	-0.36
171	the frequency of	27	13	54	26	-0.37

172	the order of	78	36	57	27	-0.37
173	the way that	124	58	31	15	-0.37
174	function of the	33	15	74	35	-0.37
175	of the two	63	29	155	73	-0.39
176	different from the	28	13	40	19	-0.39
177	the structure of	31	14	76	36	-0.42
178	what are the	130	60	27	13	-0.42
179	is that it	95	44	100	47	-0.42
180	the way in	65	30	84	40	-0.42
181	to use the	58	27	88	42	-0.44
182	be the case	32	15	35	17	-0.45
183	means that the	28	13	52	25	-0.48
184	value of the	27	13	71	34	-0.49
185	of the system	34	16	86	41	-0.50
186	of view of	28	13	39	18	-0.51
187	the work of	24	11	107	51	-0.54
188	example of a	49	23	27	13	-0.54
189	is the case	29	13	53	25	-0.55
190	is that there	49	23	40	19	-0.58
191	of the second	32	15	67	32	-0.58
192	the change in	32	15	36	17	-0.58
193	so that the	81	38	120	57	-0.59
194	is not the	58	27	118	56	-0.60
195	the area of	23	11	50	24	-0.61
196	form of the	23	11	63	30	-0.62
197	that is the	140	65	111	52	-0.63
198	and in the	116	54	180	85	-0.64
199	and the second	43	20	34	16	-0.66
200	of the fact	32	15	49	23	-0.67
201	the first is	21	10	56	26	-0.70
202	that in the	132	61	98	46	-0.77
203	and the same	36	17	35	17	-0.84
204	out that the	31	14	31	15	-0.91
205	the example of	30	14	24	11	-0.93
206	that in a	68	32	28	13	-1.08
207	is for the	34	16	24	11	-1.29

Appendix B: Written AFL Top 200 (sorted by 2-factor FTW-score)

		Speech		Writing		
		Raw freq	Freq per million	Raw freq	Freq per million	FTW
1	on the other hand	86	40	251	119	2.84
2	due to the fact that	5	2	27	13	2.64
3	on the other hand the	6	3	50	24	2.55
4	it should be noted	0	0	36	17	2.51
5	it is not possible to	1	0	31	15	2.44
6	a wide range of	9	4	66	31	2.42
7	there are a number of	11	5	30	14	2.41
8	in such a way that	20	9	23	11	2.32
9	take into account the	5	2	24	11	2.27
10	as can be seen	0	0	32	15	1.79
11	it is clear that	6	3	69	33	1.72
12	take into account	17	8	41	19	1.70
13	can be used to	11	5	95	45	1.64
14	in this paper we	0	0	29	14	1.64
15	are likely to	16	7	129	61	1.61
16	in the next section	0	0	32	15	1.60
17	a large number of	16	7	47	22	1.59
18	the united kingdom	2	1	54	25	1.57
19	on the basis of the	8	4	48	23	1.57
20	that there is no	10	5	67	32	1.56
21	over a period of	10	5	27	13	1.55
22	as a result of the	11	5	35	17	1.55
23	can be seen in	1	0	36	17	1.52
24	a wide range	13	6	69	33	1.51
25	there are a number	13	6	30	14	1.47
26	it is interesting to	0	0	32	15	1.47
27	it is impossible to	1	0	25	12	1.47
28	it is obvious that	0	0	23	11	1.46
29	it is possible to	5	2	101	48	1.46
30	it is not possible	2	1	38	18	1.45
31	been carried out	1	0	37	17	1.45
32	can be found in	0	0	39	18	1.45
33	it is important to	3	1	92	43	1.40
34	was carried out	1	0	56	26	1.39
35	is likely to be	7	3	81	38	1.38
36	wide range of	10	5	77	36	1.37
37	the same way as	10	5	32	15	1.37
38	due to the fact	5	2	27	13	1.36
39	in accordance with the	4	2	26	12	1.36
40	it is necessary to	2	1	56	26	1.35
41	the other hand	88	41	254	120	1.35

42	can be seen	12	6	185	87	1.35
43	it is likely that	0	0	39	18	1.31
44	such a way that	20	9	23	11	1.22
45	to carry out	16	7	62	29	1.22
46	it is possible that	1	0	40	19	1.22
47	with respect to the	13	6	78	37	1.20
48	give rise to	7	3	41	19	1.18
49	carried out by	4	2	43	20	1.17
50	whether or not the	6	3	38	18	1.13
51	in the present study	0	0	23	11	1.11
52	should be noted	0	0	38	18	1.07
53	be carried out	3	1	38	18	1.06
54	the other hand the	6	3	51	24	1.06
55	does not appear	3	1	27	13	1.04
56	his or her	6	3	71	34	1.01
57	is not possible to	1	0	32	15	0.99
58	shown in figure	0	0	84	40	0.96
59	be used as a	1	0	36	17	0.95
60	for the purposes of	3	1	50	24	0.95
61	be regarded as	2	1	85	40	0.94
62	to ensure that the	0	0	37	17	0.93
63	allows us to	16	7	32	15	0.93
64	it has been	26	12	168	79	0.92
65	little or no	6	3	33	16	0.90
66	carried out in	1	0	53	25	0.90
67	to distinguish between	2	1	45	21	0.88
68	in accordance with	12	6	55	26	0.88
69	they do not	13	6	118	56	0.88
70	at this stage	14	7	70	33	0.88
71	is based on the	7	3	47	22	0.88
72	shown in table	0	0	63	30	0.87
73	in the absence of	10	5	86	41	0.86
74	we have seen	11	5	56	26	0.83
75	to determine whether	4	2	33	16	0.82
76	in the context of	16	7	121	57	0.79
77	a high degree	3	1	28	13	0.78
78	the difference between the	18	8	30	14	0.78
79	an increase in the	12	6	28	13	0.78
80	it is possible	12	6	175	83	0.77
81	can be achieved	0	0	36	17	0.77
82	insight into the	0	0	34	16	0.77
83	can be expressed	3	1	49	23	0.75
84	we assume that	10	5	43	20	0.75
85	they did not	12	6	56	26	0.73
86	there has been	18	8	70	33	0.72
87	on the part of	17	8	66	31	0.70

88	in this paper	9	4	132	62	0.70
89	the purpose of this	4	2	28	13	0.70
90	less likely to	11	5	48	23	0.68
91	a large number	19	9	49	23	0.67
92	can easily be	0	0	32	15	0.67
93	with regard to	9	4	85	40	0.66
94	there are several	12	6	38	18	0.66
95	over a period	10	5	30	14	0.66
96	in this case the	17	8	57	27	0.66
97	in conjunction with	12	6	48	23	0.65
98	at the time of	14	7	68	32	0.65
99	we do not	8	4	81	38	0.64
100	has been used	8	4	43	20	0.63
101	appears to be	19	9	113	53	0.63
102	to do so	49	23	116	55	0.63
103	there are no	46	21	82	39	0.62
104	on the other	166	77	311	147	0.62
105	has also been	3	1	53	25	0.61
106	it is worth	0	0	42	20	0.61
107	can be found	2	1	69	33	0.61
108	the next section	2	1	41	19	0.60
109	are a number of	12	6	30	14	0.60
110	this paper we	0	0	34	16	0.60
111	be seen as	18	8	94	44	0.60
112	be related to the	3	1	26	12	0.59
113	to ensure that	11	5	94	44	0.59
114	it is important	6	3	139	66	0.59
115	be explained by	0	0	32	15	0.58
116	same way as	11	5	32	15	0.58
117	see for example	0	0	42	20	0.58
118	the presence of a	3	1	50	24	0.58
119	that it is not	7	3	37	17	0.58
120	in some cases	40	19	68	32	0.58
121	to the fact that	21	10	49	23	0.57
122	high levels of	12	6	35	17	0.56
123	most likely to	6	3	55	26	0.56
124	it appears that	13	6	61	29	0.56
125	it follows that	2	1	65	31	0.55
126	can also be	13	6	111	52	0.55
127	it is clear	6	3	83	39	0.54
128	by virtue of	13	6	54	25	0.54
129	the most important	46	21	112	53	0.53
130	an attempt to	25	12	62	29	0.53
131	it is impossible	2	1	36	17	0.53
132	factors such as	0	0	29	14	0.53
133	is consistent with	1	0	61	29	0.53

134	total number of	5	2	42	20	0.53
135	similar to those	0	0	47	22	0.52
136	as part of the	17	8	55	26	0.52
137	can be considered	0	0	38	18	0.52
138	at the outset	6	3	24	11	0.51
139	in more detail	7	3	27	13	0.51
140	should not be	13	6	108	51	0.51
141	could be used	9	4	41	19	0.51
142	appear to be	15	7	99	47	0.50
143	as a consequence	6	3	50	24	0.50
144	in this article	6	3	59	28	0.50
145	assumed to be	3	1	82	39	0.49
146	in the form of	19	9	98	46	0.48
147	as a whole	57	26	92	43	0.48
148	important role in	5	2	28	13	0.47
149	it is interesting	2	1	38	18	0.46
150	does not have	20	9	52	25	0.46
151	none of these	12	6	32	15	0.46
152	as shown in	1	0	139	66	0.45
153	is likely to	19	9	169	80	0.45
154	this means that	13	6	77	36	0.45
155	be noted that	0	0	45	21	0.45
156	be achieved by	0	0	28	13	0.45
157	depends on the	39	18	93	44	0.44
158	at least in	40	19	75	35	0.44
159	a small number	9	4	25	12	0.43
160	in table 1	0	0	62	29	0.43
161	in most cases	7	3	37	17	0.43
162	depending on the	30	14	62	29	0.41
163	in both cases	11	5	36	17	0.41
164	the validity of the	2	1	39	18	0.41
165	small number of	10	5	38	18	0.40
166	their ability to	16	7	40	19	0.40
167	need not be	1	0	54	25	0.40
168	needs to be	64	30	96	45	0.40
169	have shown that	4	2	63	30	0.39
170	it is necessary	5	2	71	34	0.39
171	been shown to	5	2	66	31	0.39
172	such as those	1	0	44	21	0.39
173	are as follows	1	0	34	16	0.38
174	for this purpose	3	1	31	15	0.38
175	is determined by	7	3	48	23	0.38
176	it is difficult	0	0	57	27	0.37
177	even though the	18	8	44	21	0.37
178	this does not	9	4	59	28	0.37
179	was based on	16	7	40	19	0.37

180	the nature of the	18	8	91	43	0.37
181	in the course of	28	13	58	27	0.37
182	degree to which	3	1	56	26	0.37
183	be argued that	1	0	36	17	0.36
184	in terms of a	18	8	32	15	0.36
185	for this reason	6	3	44	21	0.36
186	are based on	19	9	50	24	0.36
187	in a number of	15	7	40	19	0.36
188	two types of	14	7	45	21	0.34
189	the total number	8	4	39	18	0.34
190	is more likely	11	5	41	19	0.34
191	which can be	14	7	120	57	0.34
192	are able to	14	7	79	37	0.32
193	be considered as	0	0	46	22	0.32
194	be used to	18	8	163	77	0.31
195	b and c	11	5	37	17	0.31
196	depend on the	16	7	63	30	0.30
197	is that it is	7	3	41	19	0.30
198	is affected by	1	0	24	11	0.30
199	should also be	4	2	38	18	0.30
200	if they are	22	10	70	33	0.30

Appendix C: Spoken AFL Top 200 (sorted by 2-factor FTW-score)

		Speech		Writing		
		Raw freq	Freq per million	Raw freq	Freq per million	FTW
1	be able to	551	256	209	99	2.96
2	blah blah blah	62	29	0	0	2.92
3	this is the	732	340	127	60	2.77
4	you know what I mean	137	64	4	2	2.27
5	you can see	449	209	2	1	2.12
6	trying to figure out	41	19	2	1	2.05
7	a little bit about	101	47	0	0	2.00
8	does that make sense	63	29	0	0	1.99
9	you know what	491	228	4	2	1.99
10	the university of michigan	76	35	1	0	1.98
11	for those of you who	39	18	0	0	1.98
12	do you want me to	31	14	0	0	1.96
13	thank you very much	57	26	0	0	1.95
14	look at the	425	197	50	24	1.95
15	we're gonna talk about	42	20	0	0	1.95
16	talk a little bit	40	19	0	0	1.92
17	if you look at	173	80	0	0	1.89
18	and this is	533	248	43	20	1.87

19	if you look at the	58	27	0	0	1.80
20	no no no no	66	31	0	0	1.78
21	at the end of	191	89	128	60	1.74
22	we were talking about	49	23	0	0	1.65
23	in ann arbor	41	19	0	0	1.62
24	it turns out that	52	24	9	4	1.61
25	you need to	391	182	1	0	1.61
26	see what I'm saying	36	17	0	0	1.60
27	take a look at	67	31	3	1	1.58
28	you have a	463	215	8	4	1.57
29	might be able to	43	20	12	6	1.56
30	at the end	295	137	140	66	1.48
31	you want to	369	171	14	7	1.46
32	to do with	356	165	91	43	1.44
33	nothing to do with	48	22	19	9	1.43
34	know what I mean	140	65	7	3	1.42
35	you look at	296	137	3	1	1.42
36	university of michigan	95	44	1	0	1.42
37	what I'm talking about	29	13	0	0	1.41
38	the same thing	263	122	17	8	1.35
39	to look at	281	131	42	20	1.34
40	the end of	340	158	232	110	1.33
41	gonna be able to	38	18	0	0	1.32
42	we're talking about	132	61	0	0	1.28
43	to figure out what	26	12	2	1	1.27
44	so if you	365	170	1	0	1.24
45	so this is	373	173	0	0	1.23
46	if you want to	126	59	4	2	1.23
47	no no no	186	86	0	0	1.23
48	if you have	344	160	1	0	1.22
49	come up with a	36	17	2	1	1.21
50	we talked about	154	72	1	0	1.20
51	when you look at	47	22	1	0	1.20
52	in order to get	49	23	8	4	1.19
53	the end of the	190	88	124	59	1.19
54	oh my god	68	32	0	0	1.17
55	come up with	146	68	6	3	1.16
56	I was gonna say	56	26	0	0	1.16
57	and then you	366	170	2	1	1.16
58	a kind of	322	150	50	24	1.16
59	it doesn't matter	109	51	1	0	1.15
60	has to do with	67	31	7	3	1.14
61	you can look at	54	25	0	0	1.13
62	do you want me	34	16	0	0	1.12
63	little bit about	103	48	0	0	1.12
64	if you look	252	117	0	0	1.10

65	I just wanted to	60	28	0	0	1.10
66	you're talking about	123	57	0	0	1.08
67	what does that mean	48	22	0	0	1.08
68	the best way to	39	18	14	7	1.08
69	if you want	241	112	6	3	1.06
70	you know what i	158	73	4	2	1.05
71	we've talked about	52	24	1	0	1.05
72	we'll talk about	73	34	0	0	1.03
73	let me just	94	44	0	0	1.02
74	I was talking about	31	14	0	0	1.02
75	has to be	247	115	96	45	1.01
76	to talk about	201	93	20	9	1.00
77	it turns out	83	39	14	7	1.00
78	those of you who	58	27	1	0	0.99
79	you might want to	41	19	0	0	0.99
80	first of all	208	97	24	11	0.98
81	and so on and so	37	17	1	0	0.98
82	there was a	270	125	114	54	0.97
83	at the university of	47	22	18	8	0.97
84	yes yes yes	64	30	0	0	0.97
85	you can see that	96	45	1	0	0.96
86	I have a question	67	31	0	0	0.96
87	it has to be	80	37	13	6	0.93
88	we need to	220	102	64	30	0.92
89	what I’m saying	125	58	0	0	0.92
90	you want me to	47	22	1	0	0.92
91	all sorts of	107	50	2	1	0.91
92	as you can see	44	20	0	0	0.90
93	to figure out	114	53	8	4	0.90
94	keep in mind	47	22	6	3	0.90
95	what do you mean	63	29	1	0	0.89
96	it looks like	143	66	2	1	0.88
97	let's look at	82	38	0	0	0.87
98	you look at the	89	41	0	0	0.87
99	to make sure	123	57	13	6	0.86
100	if you wanted to	41	19	0	0	0.85
101	make sure that	121	56	15	7	0.84
102	end up with	81	38	9	4	0.84
103	and you can see	85	39	0	0	0.84
104	came up with	67	31	2	1	0.84
105	doesn't have to be	36	17	0	0	0.83
106	I mean if you	88	41	0	0	0.83
107	you've got a	124	58	0	0	0.83
108	gonna talk about	89	41	0	0	0.82
109	how many of you	37	17	0	0	0.82
110	I mean if	223	104	0	0	0.81

111	look at it	173	80	5	2	0.81
112	piece of paper	34	16	5	2	0.81
113	and so forth	129	60	35	17	0.80
114	and you can	306	142	6	3	0.79
115	looking at the	180	84	25	12	0.79
116	we're gonna talk	49	23	0	0	0.79
117	go back to the	48	22	9	4	0.79
118	you know what I'm	52	24	0	0	0.76
119	that you can	292	136	2	1	0.76
120	we're looking at	56	26	0	0	0.76
121	what I mean	219	102	12	6	0.74
122	do you know what	67	31	2	1	0.73
123	how do you know	42	20	4	2	0.73
124	you don't need to	42	20	2	1	0.73
125	you're looking at	68	32	0	0	0.72
126	turns out that	61	28	9	4	0.72
127	it could be	180	84	48	23	0.72
128	figure out what	56	26	2	1	0.72
129	if you've got	69	32	0	0	0.72
130	I wanted to	180	84	6	3	0.71
131	you could you could	33	15	0	0	0.71
132	might be able	44	20	12	6	0.70
133	trying to figure	44	20	2	1	0.70
134	what you're saying	86	40	1	0	0.67
135	we have to	252	117	43	20	0.67
136	I'm talking about	68	32	0	0	0.67
137	so you can	245	114	1	0	0.66
138	this kind of	205	95	49	23	0.65
139	don't worry about	29	13	0	0	0.65
140	it's gonna be	151	70	0	0	0.64
141	if you have a	96	45	0	0	0.64
142	wanna talk about	45	21	0	0	0.64
143	so you can see	38	18	0	0	0.64
144	I want you to	79	37	0	0	0.63
145	to look at the	59	27	15	7	0.63
146	to each other	98	46	50	24	0.62
147	the kind of	257	119	50	24	0.62
148	at this point	116	54	66	31	0.61
149	one of these	189	88	50	24	0.60
150	and if you	284	132	4	2	0.60
151	you think about it	55	26	0	0	0.59
152	talk about the	160	74	5	2	0.59
153	it might be	138	64	76	36	0.59
154	for those of you	49	23	0	0	0.59
155	to do with the	93	43	39	18	0.59
156	I'm not gonna	97	45	0	0	0.58

157	was talking about	82	38	1	0	0.58
158	have to do with	42	20	5	2	0.58
159	tell me what	53	25	2	1	0.57
160	look at this	123	57	3	1	0.57
161	in a sense	160	74	32	15	0.56
162	okay I don't know	31	14	0	0	0.56
163	I'll talk about	31	14	0	0	0.56
164	you need to do	33	15	0	0	0.56
165	do you want	149	69	6	3	0.55
166	we talk about	89	41	1	0	0.54
167	any questions about	31	14	0	0	0.53
168	come back to	79	37	3	1	0.53
169	you can see the	61	28	0	0	0.53
170	the reason why	78	36	16	8	0.52
171	it in terms of	31	14	4	2	0.52
172	what I want to	37	17	6	3	0.52
173	we looked at	48	22	6	3	0.51
174	if you wanna	138	64	0	0	0.51
175	take a look	89	41	3	1	0.50
176	if you were to	47	22	0	0	0.50
177	I'll show you	45	21	0	0	0.49
178	talking about the	137	64	6	3	0.49
179	that make sense	67	31	2	1	0.49
180	this is this is	84	39	0	0	0.48
181	how do we	126	59	10	5	0.48
182	we were talking	55	26	1	0	0.48
183	wanna look at	41	19	0	0	0.48
184	you're trying to	81	38	0	0	0.47
185	a look at	131	61	10	5	0.47
186	if you were	163	76	7	3	0.47
187	you're interested in	46	21	0	0	0.47
188	to think about	175	81	11	5	0.46
189	gonna be able	38	18	0	0	0.46
190	by the way	141	65	9	4	0.45
191	we look at	93	43	15	7	0.45
192	I think this is	57	26	1	0	0.45
193	but if you	203	94	5	2	0.45
194	at some point	51	24	15	7	0.44
195	I'm gonna go	51	24	0	0	0.44
196	thank you very	59	27	1	0	0.43
197	can look at	74	34	1	0	0.43
198	what happens is	86	40	0	0	0.43
199	on the board	65	30	6	3	0.42
200	um let me	37	17	0	0	0.42

[1] MICASE speech events include lectures, seminars, student presentations, office hours, and study groups; for further details about the specific genres in MICASE, see Simpson-Vlach and Leicher (2006). BNC spoken academic files include primarily lectures and tutorials. BNC written academic texts include research articles and textbooks.

[2] Furthermore, this was the only corpus of conversational American English speech available to us; although telephone conversations are not necessarily ideal, they were quite adequate for comparison purposes.

[3] Because these formulas appeared frequently in *both* spoken and written genres, the minimum threshold was set at six out of nine of the disciplinary sub-corpora, which had to include both written and spoken corpora. In fact, over 100 of the Core AFL formulas appeared in at least eight out of nine, and furthermore most of them occurred at frequencies well over 20 times per million.