An Academic Formulas List (AFL): Corpus Linguistics, Psycholinguistics, and Education

Nick C. Ellis & Rita Simpson-Vlach
University of Michigan

San José State University

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. We assess the instructional and psycholinguistic validity of these formulas in order to prioritize them 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), Corpus Linguistics, Psycholinguistics, Education, English for Academic Purposes

Introduction

The specific aim in this research is to create an Academic Formulas List (AFL), a pedagogically useful list of formulaic sequences for academic speech and writing comparable to the Academic Word List (hereafter AWL, Coxhead 2000). Corpus linguistic analyses of written and spoken academic discourse allow us to identify recurring, high-frequency lexical bundles, phrases, or formulas, and research has shown that these are important characteristics of academic registers (Biber, Conrad et al. 2004; Simpson 2004). Cognitive scientific analyses also inform us that knowledge of these formulas is crucial for fluent processing. Second language acquisition researchers and EAP practitioners need a prioritized list of the most important formulas characterizing academic discourse, which as of yet has not been available.

Our research therefore triangulates the construct of 'formula' from corpus linguistic, psycholinguistic and educational perspectives.

Corpus extraction of the AFL

Three, four, and five word formulas occurring at least 10 times per million words were extracted from corpora of 2.1M words of academic spoken language from MICASE (Simpson, Briggs, Ovens, & Swales, 2002) and selected academic spoken BNC files (British National Corpus, 2006), 2.1M words of academic written language from Hyland's (2004) research article corpus, plus selected academic writing BNC files, 2.9M words of non-academic speech from the Switchboard (2006) corpus, and 1.9M words of non-academic writing from the FLOB and Frown corpora gathered in 1991 to reflect British and American English over 15 genres (ICAME, 2006).

The program Collocate (Barlow, 2004) allowed us to measure the frequency of each n-gram along with the mutual information (MI) score for each phrase. 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 often than would be expected by chance; it is a measure of how much they cohere or are found in collocation. 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. 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. High MI n-grams, in contrast, are those with much greater coherence than is expected by chance, and this tends to correspond with distinctive function or meaning as well as grammatical well-formedness as a complete phrase.

The total number of formulas appearing in any one of the four corpora at the threshold level of 10 per million was approximately 14,000. In order to determine which formulas were more frequent in the academic corpora than in their non-academic counterparts, we used the log-likelihood (LL) statistic to identify the formulas which were statistically more frequent, at a significance level of $p<.01$, in the academic corpora than in their non-academic counterparts. We separately compared academic speech vs. non-academic speech, resulting in over 2000 items, and academic writing vs. non-academic writing resulting in just under 2000 items.

Instructional Validation of Academic Formulas

Our investigation of educational validity of these academic formulas used a representative sample of 108 of them, 54 from the Speech list 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. Example items are shown in Table 1.

Table 1

Sample formulaic sequences factorially crossing *n*-gram Length (3, 4, 5), Frequency (low, medium, high), and Mutual Information (low, medium, high)

		Mutual Information		
		Low (3.3)	Medium (6.7)	High (11)
Frequency (n per million)	Low (10.9)	that the only the length of the in the context of the	happens is that and so on but as in the case of	circumstances in which it has been shown of the court of appeal
	Medium (15.0)	and at the the value of the the way in which the	that may be the relationship between the it is not possible to	see for example a wide variety of it should be noted that
	High (43.6)	the content of is one of the in the case of the	a kind of the extent to which at the beginning of	in other words a great deal of it can be seen that

The stratified sample of 108 *n*-grams in total constituted the stimuli for the Instructor judgments of formulaicity and the Psycholinguistic Processing experiments. We asked 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 6 raters with an inter-rater $\alpha = 0.77$.

B. whether or not they thought the phrase had ‘a cohesive meaning or function, as a phrase’. There were 8 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 6 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 intercorrelation reassured us as to the reliability and validity of these instructor insights. We then investigated which of Frequency or MI better predicted the insights. Correlation analysis suggested that while both of these dimensions contributed to instructors valuing the formula, it was MI which most 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.

The high intercorrelations of the instructor ratings suggest a latent factor of formulaicity underlying their judgments. The significant associations between the corpus metrics of *n*-gram frequency and MI, and the various instructor judgments of *n*-gram formulaicity, identifiably of function, and teaching-worth suggest a successful triangulation of instructor insights and corpus metrics: In other words, these corpus-derived measures do serve to identify *n*-grams that instructors judge to be clearly identifiable formulas which are worth teaching. Both *n*-gram frequency and MI factor into this prediction, but it is the MI of the string – the degree to which the words are bound together – that is the major determinant.

Psycholinguistic Validation of Academic Formulas

A representative sample of these was taken which factorially combined Frequency (high, medium, and low frequency of occurrence), Mutual Information (high, medium, and low MI, a statistical measure of the degree to which the words in the phrase are associated more than would be expected by chance), Length (3, 4, 5 word), and Source (spoken or written language). Four experiments then determined which of these factors affected the accuracy and fluency of processing of these formulas in native language speakers of English and in advanced second language learners of English (all students at a large North American University). The language processing tasks were: (1) rate of reading and rate of spoken articulation, (2) speed of reading and acceptance in a grammaticality judgment task where half of the items were real phrases in English and half were not, (3) speed of comprehension and acceptance of the formula as being appropriate in a sentence context, (4) binding and primed pronunciation: the degree to which reading the beginning of the formula primed recognition of its final word. These tasks were selected to sample an ecologically valid range of language processing skills: spoken and written, production and comprehension, form-focused and meaning-focused. Processing in all experiments was affected by the various corpus-derived measures: length, frequency, MI, and source, but to very different degrees in the different learners. For native speakers it is predominantly the MI of the formula which determines its processability. For non-native learners of the language it is predominantly the frequency of the formula which determines its accuracy and fluency of processing. These findings have important implications for the psycholinguistic validity of corpus-derived formulas, their acquisition, and their instruction.

The Academic Formulas List (AFL)

The resultant 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 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.

Sub-section heading (not numbered, Times 12, left alignment)

Write paper in Word or rtf format. Use 1.5 line spacing throughout, except for abstract and bio note at the end, which should be in single line spacing. Set spacing before and after paragraphs to zero. Set all margins (top, bottom left and right) to 3 cm.

Title, author's name, affiliation and country, abstract, keywords, section and sub-section headings, body text, captions and references should use the same fonts and font sizes as this stylesheet. Skip a single line in between paragraphs and do not use paragraph indentations. Concordances, pictures, graphs and figures should be included in the text at approximate point of insertion, with captions underneath.

on request. No reception is possible but **sample** newspaper files can be accessed. [p] [c]
for your first recipes right now. Then you can **sample** the delights of `SIMPLY DELICIOUS'' in your
modem. And a telephone line, of course. For a **sample** of what is actually available today, the best
ids to CD track increment flags, with built in **sample** rate conversion for the audio. A built-in
[h] Disney delights [/h] There's still time to **sample** the thrills and spills of Euro Disney this
looking at descriptive characteristics of the **sample** (as assessed from a respondent information
Truth Dog', with its eerie Neville Chamberlain **sample** and caustic guitar scraping, sounds like PWEI
song I play before I go out on weekends. The **sample** is from `Dazz' by Brick, a big hit from the
can be added by loading from ROM cards, MIDI **sample** dumps or Yamaha tx

Caption (not numbered, Arial 10, regular, separated by commas; skip three lines at the end)

References to other publications in body text should contain author’s surname and date or author's name, date, colon and page number in parentheses, e.g. (Smith 2005) and (Jones 1958: 11). Full bibliographical references (only those cited in text) should be listed in alphabetical order at the end of the paper and follow the examples given at the end of this stylesheet.

Name the file with the first author's surname followed by Talc8 (e.g., SmithTalc8.doc). Save the paper in doc or rtf format and send it as an email attachment to talc8lisbon@gmail.com. Write "proceedings paper" in the subject line. The deadline for submitting your paper is 30 May 2008.

References

Barlow, M. 2004. *Collocate*. Houston: Athestan Publications.

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

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

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

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

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

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., Briggs, S., Ovens, J., & Swales, J. M. 2002. *The Michigan Corpus of Academic Spoken English*. [Electronic Version]. The Regents of the University of Michigan from http://www.hti.umich.edu/m/micase.

Switchboard. 2006, August 5, 2006. *A User's Manual*. from http://www.ldc.upenn.edu/Catalog/docs/switchboard/

The authors

Nick Ellis is a Research Scientist and Professor of Psychology at the University of Michigan. His research interests include language acquisition, cognition, reading in different languages, corpus linguistics, cognitive linguistics, and applied psycholinguistics. Currently his research focuses on second language acquisition, particularly (1) explicit and implicit language learning and their interface, (2) usage-based acquisition and the probabilistic tuning of the system, (3) vocabulary and phraseology, (4) language and brain, (5) the advanced language learner, (6) applications of psychological theory in language testing and instruction, (7) learned attention and language transfer, (8) emergentist accounts of language acquisition.

Rita Simpson-Vlach was a Research Associate at the English Language Institute of the University of Michigan, where she served as project director of the Michigan Corpus of Academic Spoken English from its inception until 2006. Most recently she has been a lecturer in the Department of Linguistics and Language Development at San José State University. Her research interests lie mainly in the area of corpus linguistics and EAP, specifically in the use of corpora for pragmatic and discourse analyses and for use in EAP teaching materials development.