Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL

NICK C. ELLIS
University of Michigan
Ann Arbor, Michigan, United States

RITA SIMPSON-VLACH
San José State University
San José, California, United States

CARSON MAYNARD
University of Michigan
Ann Arbor, Michigan, United States

Natural language makes considerable use of recurrent formulaic patterns of words. This article triangulates the construct of *formula* from corpus linguistic, psycholinguistic, and educational perspectives. It describes the corpus linguistic extraction of pedagogically useful formulaic sequences for academic speech and writing. It determines English as a second language (ESL) and English for academic purposes (EAP) instructors' evaluations of their pedagogical importance. It summarizes three experiments which show that different aspects of formulaicity affect the accuracy and fluency of processing of these formulas in native speakers and in advanced L2 learners of English. The language processing tasks were selected to sample an ecologically valid range of language processing skills: spoken and written, production and comprehension. Processing in all experiments was affected by various corpus-derived metrics: length, frequency, and mutual information (MI), but to different degrees in the different populations. For native speakers, it is predominantly the MI of the formula which determines processability; for nonnative learners of the language, it is predominantly the frequency of the formula. The implications of these findings are discussed for (a) the psycholinguistic validity of corpus-derived formulas, (b) a model of their acquisition, (c) ESL and EAP instruction and the prioritization of which formulas to teach.

Corpus linguistic research demonstrates that natural language makes considerable use of recurrent multiword patterns or *formulas* (Ellis, 1996, 2008a; Granger & Meunier, in press; Pawley & Syder, 1983; Sinclair, 1991, 2004; Wray, 2002). Sinclair (1991) summarized the results of

corpus investigations of such distributional regularities: "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" (p. 100), and suggested that for normal texts, the first mode of analysis to be applied is the idiom principle, as most text is interpretable by this principle. Erman and Warren (2000) estimate that about half of fluent native text is constructed according to the idiom principle. Comparisons of written and spoken corpora suggest that formulas are even more frequent in spoken language (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Brazil, 1995; Leech, 2000). English utterances are constructed as intonation units that have a modal length of four words (Chafe, 1994) and that are often highly predictable in terms of their lexical concordance (Hopper, 1998). Speech is constructed in real time and this imposes greater working memory demands compared with writing, hence the greater need to rely on formulas: It is easier for us to look something up from long-term memory than to compute it (Bresnan, 1999; Kuiper, 1996).

Psycholinguistic research demonstrates language users' sensitivity to the frequencies of occurrence of a wide range of different linguistic constructions (Ellis, 1996, 2002a, 2002b, 2008c) and therefore provides clear testament of the influence of each usage event, and the processing of its component constructions, on the learner's system. Usage-based theories of language consequently analyze how frequency and repetition affect, and ultimately bring about, form in language, and how this knowledge affects language comprehension and production (Bod, Hay, & Jannedy, 2003; Bybee & Hopper, 2001; Ellis, 2002b, 2008b; Hoey, 2005; Robinson & Ellis, 2008).

Research in this area has produced evidence that language processing is sensitive to formulaicity and collocation. For formulaicity, Swinney and Cutler (1979) found that study participants took much less time to judge idiomatic expressions, such as *kick the bucket*, as being meaningful English phrases than they did for nonidiomatic control strings like *lift the bucket* (see also Conklin & Schmitt, 2007; Schmitt, 2004). For collocation, Ellis, Frey, and Jalkanen (in press) used lexical decision tasks to demonstrate that native speakers preferentially recognized frequent verb-argument and booster/maximizer-adjective pairs than they did less frequent ones. McDonald and Shillcock (2004) used eye movement recording to reveal that the reading times of individual words are affected by the transitional probabilities of the lexical components. So with sentences like *One way to avoid confusion/discovery is to make the changes during the vacation*, readers read high transitional probability sequences such as *avoid confusion* faster than low transitional probability like *avoid discovery*. Jurafsky, Bell, Gregory, and Raymond (2001) analyzed the articulation time of successive two-word sequences in the SwitchBoard corpus (University of Pennsyl-

vania Linguistic Data Corpus, n.d.) to show that in production, humans shorten words that have a higher contextualized probability. This phenomenon is entirely graded, with the degree of reduction a continuous function of the frequency of the target word and the conditional probability of the target given the previous word. The researchers argue on the basis of this evidence that the human production grammar must store probabilistic relations between words. As Bybee (2003) quips, on a variant of Hebb's (1949) learning rule later encapsulated in the paraphrase "Cells that fire together, wire together," "Items that are used together fuse together."

These experiments demonstrate sensitivity to formulaicity in native fluent speakers, but we have yet to discover the psycholinguistic and corpus linguistic determinants of this sensitivity, and to compare these effects in second language learners and native speakers. There is considerable interest in formulaic language in second language acquisition (SLA), as recent reviews attest (Cowie, 2001; Gries & Wulff, 2005; Meunier & Granger, 2008; Robinson & Ellis, 2008; Schmitt, 2004; Wray, 2002). English for academic purposes (EAP) research (e.g., Flowerdew & Peacock, 2001; Hyland, 2004; Swales, 1990) focuses on determining the functional patterns and constructions of different academic genres. Every genre has a characteristic form of expression, and learning to be effective in the genre involves mastering this phraseology. So lexicographers, guided by representative corpora (Hunston & Francis, 1996; Ooi, 1998), develop learner dictionaries which focus on examples of usage as much as, or even more than, on definitions. Corpora now play central roles in identifying relevant constructions for language teaching (Cobb, 2007; Römer, in press; Sinclair, 1996). Large samples of writing or speech such as the Michigan Corpus of Academic Spoken English (MICASE; English Language Institute of the University of Michigan, 2002) are assembled in ways that adequately represent different academic fields and registers; linguists, then, engage in qualitative investigation of patterns, at times supported by computer software for the analysis of concordances and collocations.

Analyses of such academic corpora demonstrate that academic discourse contains a high frequency of common lexical bundles such as *in order to*, *the number of*, *the fact that*, *as __ as __*, (Biber, Conrad, & Cortes, 2004), collocations and formulaic sequences such as *research project*, *as a result of*, *to what extent*, *in other words* (Schmitt, 2004; Simpson-Vlach & Ellis, in press), and idioms such as *come into play*, *bottom line*, *rule of thumb*, *ball-park estimate* (Simpson & Mendis, 2003). The learner has to know these idioms as a whole; a literal interpretation is no good. And they have to know the common collocations and lexical bundles, too, not only to increase their reading speed and comprehension (Grabe & Stoller, 2002), but also to be able to write in a nativelike fashion: It is not enough

to know the meaning of words like *describe* or *advantage* or *mistake* if the language user doesn't know how to use them and writes "describe about the problem" rather than "describe the problem," "get advantage of" rather than "take advantage of," or "did the mistake" rather than "made the mistake." Even advanced language learners have considerable difficulty with collocations, often resulting from transfer of first language (L1) combinatorial restrictions, and the frequency of these problems shows that learners need instruction in these aspects of language (Nesselhauf, 2003).

Thus, despite formulas being one of the hallmarks of child second language development (McLaughlin, 1995) and, as the American Council on the Teaching of Foreign Languages (ACTFL, 1999) guidelines demonstrate, their being central in novice adult learners' second language, too (Ellis, 1996, 2003), advanced learners of second language have great difficulty with nativelike collocation and idiomaticity. Many grammatical sentences generated by language learners sound unnatural and foreign (Granger, 1998; Howarth, 1998; Pawley & Syder, 1983). This dissociation with proficiency suggests that the formulaic knowledge of the novice is different from that of the fluent language user and is created differently.

The difficulty second language learners have in attaining nativelike formulaic idiomaticity and fluency raises issues of instruction (Meunier & Granger, 2008; Schmitt, 2004). Within the language learning and teaching literature, Nattinger and DeCarrico (1992) argue for the *lexical phrase* as the pedagogically applicable unit of prefabricated language. Nattinger (1980) argues that

> for a great deal of the time anyway, language production consists of piecing together the ready-made units appropriate for a particular situation and . . . comprehension relies on knowing which of these patterns to predict in these situations. Our teaching therefore would center on these patterns and the ways they can be pieced together, along with the ways they vary and the situations in which they occur. (p. 341)

The *lexical approach* (Lewis, 1993), similarly predicated on the idiom principle, focuses instruction on relatively fixed expressions that occur frequently in spoken language.

In sum, the pervasive nature of formulaic language has a number of important consequences for TESOL. English language researchers and practitioners need

- to identify those formulas that have high utility for language learners.
- to develop an understanding of how best to integrate formulaic language into the learning curriculum, and how best to instruct learners in its use.

- a clearer understanding of the psycholinguistics of formulaic language in native speakers and second language learners and of the factors that determine learnability and processing fluency.
- to let these understandings inform which formulas should be prioritized for instruction in learners at different stages of development and need.

The current article summarizes some of our research into these areas. The available article length does not allow us to give much detail, and the reader is referred to other instances of our work (Simpson-Vlach & Ellis, in press) for a fuller description of our methods, the resulting list of academic formulas, their functional classification, and their prioritization.

To contextualize our interests, as an English language institute at a major U.S. university with a high proportion of international graduate students studying in English as the language of instruction, our goal is to create an empirically derived and pedagogically useful list of formulaic sequences for academic speech and writing, an Academic Formulas List (AFL) comparable to the Academic Word List (Coxhead, 2000). We are motivated by current developments in language education, corpus linguistics, cognitive science, SLA, and EAP. Research and practice in second language education demonstrates that academic study puts substantial demands on students because the relevant 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 formulaic expressions are not only salient but also functionally significant: Cognitive science demonstrates that knowledge of these formulas is crucial for fluent processing. And current trends in SLA and EAP demand ecologically valid instruction that identifies and prioritizes the most important formulas in different genres.

IDENTIFYING RELEVANT FORMULAIC EXPRESSIONS

We used corpus linguistic techniques to identify the academic formulas in corpora of written and spoken discourse that are significantly more common in academic discourse than in nonacademic discourse and which occupy a range of academic genres or habitats. Three-, four-, and five-word formulas occurring at least 10 times per million words were extracted from corpora of 2.1 million words of academic spoken language from MICASE (English Language Institute of the University of Michigan, 2002) and selected academic spoken language files from the

British National Corpus (BNC; BNC Consortium, 2006), 2.1 million words of academic written language from Hyland's (2004) research article corpus, plus selected academic writing files from the BNC, 2.9 million words of nonacademic speech from the Switchboard corpus (University of Pennsylvania Linguistic Data Consortium, n.d.), and 1.9 million words of nonacademic writing from the Freiburg Lancaster Oslo/Bergen (FLOB) and Frown corpora gathered in 1991 to reflect British and American English over 15 genres (ICAME, 1999).

The software program Collocate (Barlow, 2004) allowed us to measure the frequency of each *n*-gram along with the 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 (Manning & 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. 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 coherence 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. We used the log-likelihood (LL) statistic (Oakes, 1998) to identify the formulas which were statistically more frequent, at a significance level of $p < 0.01$, in the academic corpora than in their nonacademic counterparts. We separately compared academic speech versus nonacademic speech, resulting in over 2,000 items, and academic writing versus nonacademic writing, resulting in just under 2,000 items.

THE INSTRUCTIONAL VALUE OF THE 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 writing 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, Frequency, and Mutual Information

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

Note. 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*):

1. whether they thought the phrase constituted a formulaic expression, or fixed phrase, or chunk. There were 6 raters with an interrater $\alpha = 0.77$.
2. whether they thought the phrase had a cohesive meaning or function, as a phrase. There were 8 raters with an interrater $\alpha = 0.67$.
3. whether they thought the phrase was worth teaching, as a bona fide phrase or expression. There were 6 raters with an interrater $\alpha = 0.83$.

Formulas which scored high on one of these measures tended to score high on another: r AB = 0.80, $p < 0.01$; r AC = 0.67, $p < 0.01$; r BC = 0.80, $p < 0.01$). 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 whether frequency or MI better predicted the insights. Correlation analysis suggested that although both of these dimensions contributed to instructors valuing the formula, it was MI which most influenced their prioritization: r Frequency/A = 0.22, $p < 0.05$; r Frequency/B = 0.25, $p < 0.05$; r Frequency/C = 0.26, $p < 0.01$; r MI/A = 0.43, $p < 0.01$; r MI/B = 0.51, $p < 0.01$; r MI/C = 0.54, $p < 0.01$. 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, identifiability of function, and teaching-worthiness 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.

THE PSYCHOLINGUISTIC VALIDITY OF THE FORMULAS IN NATIVE AND ESL SPEAKERS

We used the same 108-item subset to investigate the psycholinguistic aspects of these formulas in three different experiments. The items in the subset were selected to sample an ecologically valid range of language processing skills—spoken and written, production and comprehension—while permitting rigorous measurement of processing. The language processes investigated were (a) speed of reading and acceptance in a grammaticality judgment task where half of the items were real phrases in English and half were not, (b) rate of reading and rate of spoken articulation, and (c) binding and primed articulation—the degree to which reading the beginning of the formula primed recognition of its final word.

Experiment 1. Reading and Recognition in a Grammaticality Judgment Task

Method

Participants were asked to judge whether visually presented word strings were likely to be found in English or not. The instructions were

> On each trial we will show you a string of words and we want you to judge whether you think you are likely to come upon such a sequence in English. For example, you might read or hear such strings as 'in the road', 'open your books to', 'where are the', but you would not read or hear 'on phone the', 'by way the', 'put on shirt his'. You begin each trial by pressing the space bar. A string is shown mid screen. If you think it's English, press 'yes', if you are not likely to read or hear this in English, press 'no'. We are measuring how quickly and how correctly you do this.

The 108 real phrases and 108 nonphrases made by scrambling the word orders of formulas were randomly ordered. The experiment was run on Dell computers under Microsoft Windows XP using E-Prime 1.1 (Psychology Software Tools, 2002). Responses were measured using the E-Prime Serial Response box. Note was taken both of the correctness of participants' responses and their reaction times (RTs). Outliers, defined as responses less than 200 milliseconds (ms) or more than 3 standard deviations above the participant's mean were replaced by the mean value for that participant. RTs for correct *yes* responses on the 108 real formulas were averaged across participants and analyzed using multiple regression seeking the effects of word length, frequency, and MI.

Participants

The native speaker group comprised 11 students or staff from the University of Michigan whose first language was English. There were 7 females and 4 males. Their ages ranged from 18 to 33, average 23.4 years.

The ESL group were 11 international students at the University of Michigan taking EAP classes at the English Language Institute. Their first languages were Chinese (5), Thai (4), Korean (1), and Spanish (1). There were 6 females and 5 males. Their ages ranged from 21 to 46, average 31.3 years. Their English language proficiency was sufficient to permit enrollment at the university for a graduate degree through the medium of English. They had studied English for between 10 and 30 years, average 15.1. They had been immersed in English-medium studies at the university for between 1 and 30 months, average 8.1. All participants were paid US$10 for taking part.

Results

Accuracy of responding was greater than 96%. The interparticipant reliability of RT responses was $\alpha = 0.68$.

For the native speakers, a forced entry multiple regression predicted RT from word length, frequency, and MI as the independent variables. It showed significant effects of *n*-gram length ($\beta = 0.71$—the more words in the formula, the longer the judgment time) and of MI ($\beta = -0.52$—the greater the coherence of the formula, the shorter its judgment time), but not of frequency. These data are detailed in Table 2.

The same analysis for the advanced ESL learners also showed significant effects of *n*-gram length ($\beta = 0.38$) but unlike the native speakers, ESL learner judgment time was significantly associated with the frequency of the formula in the input ($\beta = -0.24$), rather than its MI.

TABLE 2

Multiple Regressions Predicting Reading Recognition Reaction Times (RTs) in Native Speakers and Advanced ESL Learners in Experiment 1

Dependent variable judgment RT	Predictors			R^2
	n-gram length	Frequency	Mutual information	
Native English speakers				
β	0.71	–0.04	–0.52	19%
p	0.001***	n.s.	0.001***	
Advanced ESL learners				
β	0.38	–0.24	–0.07	20%
p	0.012*	0.009**	n.s.	

Note. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

Discussion

The fact that recognition of the formulas was affected by these factors provides evidence for the psycholinguistic reality of formulaicity as defined and derived by corpus linguistic means. It is notable, however, that native speakers and ESL learners are sensitive to different metrics: For native speakers, like the instructors who were judging these strings for different aspects of formulaicity in the previous section, it is the MI of the string, the degree to which the words cohere at levels above those expected by chance, that influences their processing. In contrast, formula processing in the nonnatives, despite their many years of ESL instruction, was a result of the frequency of the string rather than its coherence. For learners at this stage of development, it is the number of times the string appears in the input that determines fluency. We will return to these differences in due course, suggesting a model of acquisition which might explain them.

Experiment 2. Reading Aloud: Voice Onset and Articulation Time

Methods

Participants were shown the formulaic strings one at a time on a computer screen and instructed to read them aloud as quickly as possible. The experiment was run on Dell computers under Microsoft Windows XP using E-Prime 1.1 (Psychology Software Tools, 2002). The beginning of each new string on the monitor was accompanied by a short beep. We audio recorded each session and later analyzed the recordings using Praat (Boersma & Weenink, 2007). For each trial, we measured the

pause between the onset of the written string and the beginning of the participant's spoken response. This will be referred to as VOT (voice onset time). Outliers were dealt with as in Experiment 1. We also analyzed *articulation time*—the duration between the participant's speech onset and offset. VOT thus measures the time the participant takes to read the formula and assemble a pronunciation for it. Articulation time measures the time taken to utter the string. The VOTs and articulation times were averaged across participants and analyzed using multiple regression, looking for the effects of word length, frequency, and MI.

Participants

The native speaker group comprised 6 students or staff from the University of Michigan whose first language was English. There were 4 females and 2 males. Their ages ranged from 19 to 21, average 20.0 years.

The ESL group were 6 international students at the University of Michigan taking EAP classes at the English Language Institute. Their first languages were Chinese (4) and Korean (2). There were 3 females and 3 males. Their ages ranged from 21 to 38, average 21.2 years. Their English language proficiency was sufficient to permit their study at the university for a graduate degree through the medium of English. They had studied English for between 6 and 25 years, average 13.2. They had been immersed in English medium studies at the university for between 3 and 31 months, average 12.0. All participants were paid US$10 for taking part.

Results

For the native speakers, a forced entry multiple regression predicted VOT from formula length in words, length in spoken phonemes, frequency, and MI as the independent variables. It showed significant effects of *n*-gram length ($\beta = 0.37$), number of phonemes ($\beta = 0.25$), and MI ($\beta = -0.43$), but not of frequency. These data are detailed in Table 3. The same analysis for the advanced ESL learners showed significant effects of number of phonemes ($\beta = 0.34$), but unlike the native speakers, ESL VOT was significantly associated with the frequency of the formula in the input ($\beta = -0.20$), rather than with its MI. These data clearly parallel those for formula recognition time in Experiment 1.

For the native speakers, a forced entry multiple regression predicted articulation time from formula length in words, length in spoken phonemes, frequency, and MI as the independent variables. It showed very large significant effects of *n*-gram length ($\beta = 0.80$), but, as can be seen in Table 3, nothing else. The same analysis for the advanced ESL learners also showed significant effects of number of phonemes ($\beta = 0.75$). As

TABLE 3

Multiple Regressions Predicting Articulation Latency and Articulation Time in Native Speakers and Advanced ESL Learners in Experiment 2

Articulation Latency (Voice Onset Time)					
Dependent variable	Predictors				
Articulation latency	*n*-gram length	n.phonemes	Frequency	MI	R^2
Native English speakers					
β	0.37	0.25	–0.06	–0.43	16%
p	0.05*	0.07 ?	n.s.	0.01**	
Advanced ESL learners					
β	0.02	0.34	–0.20	–0.12	16%
p	n.s.	0.01*	0.04*	n.s.	
Articulation Time					
Dependent variable	Predictors				
Articulation time	*n*-gram length	n.phonemes	Frequency	MI	R^2
Native English speakers					
β	0.04	0.80	0.04	–0.02	65%
p	n.s.	0.001***	n.s.	n.s.	
Advanced ESL learners					
β	–0.11	0.75	–0.10	0.16	69%
p	n.s.	0.001***	0.08 ?	n.s.	

Note. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, ? marginally significant.

with the VOT analyses, there was no effect of MI but a marginal effect of formula frequency—the higher the frequency, the faster the articulation (β = –0.10).

Discussion

Both the decoding and the articulation of the formulas were shown to be affected by the corpus-derived metrics, strengthening the evidence for the psycholinguistic validity of corpus-defined formulaicity. As in Experiment 1, native speakers and ESL learners were sensitive to different metrics. For the native speakers, it was the MI of the string, the degree to which the words cohere at levels above those expected by chance, that influenced their VOT; whereas for the nonnatives it was the frequency of the string. As in Experiment 1, for learners at this stage of development, it was the number of times the string appears in the input that determined fluency of reading and the speed of assembling the motor instructions for articulation.

The lack of significant prediction of articulation execution time from formula frequency or MI in native speakers fails to support the findings of Jurafsky et al. (2001), described earlier, which indicated that the articulation time of successive two-word sequences in the SwitchBoard

corpus was shorter for sequences with a higher MI. However, in parallel with their VOT patterns, the advanced ESL learners tended to pronounce more fluently the formulas of higher input frequency.

Experiment 3: Priming of the Final Word of the Formula

Method

Participants were asked to read aloud the final word of each string as quickly as possible. They were instructed:

> Every trial starts with a fixation point (+). When you are focused on the fixation point and ready, you will press the space bar. Then, you will see either an incomplete phrase or a series of x's. Next, you will see a single word. As soon as you see this word, say it clearly into the microphone.

The experiment was run on Dell computers under Microsoft Windows XP using E-Prime 1.1 (Psychology Software Tools, 2002). Responses were measured using the E-Prime Serial Response box. On each trial, the words constituting the beginning part of the formula were presented midscreen for 2000 ms. There was then a 1000 ms. blank screen interstimulus interval before the final target word appeared in a different color. We measured the VOT between the onset of the final word of the formula and the beginning of the participant's spoken response using a microphone and voice key. Outliers were dealt with as in Experiment 1. The experiment thus measured the degree to which accessing the articulation of the final word of the formula is *primed* (made faster) by seeing the beginning part of the formula. The VOTs for the final word of the formulas were averaged across participants and analyzed using multiple regression, looking for the effects of word length, frequency, and MI.

Participants

The native speaker group comprised 18 students or staff from the University of Michigan whose first language was English. There were 11 females and 7 males. Their ages ranged from 19 to 33, average 22.3 years.

The ESL group were 16 international students at the University of Michigan taking EAP classes at the English Language Institute. Their first languages were Chinese (9), Japanese (6), and Korean (1). There were 7 females and 9 males. Their ages ranged from 19 to 34, average 24.9 years. Their English language proficiency was sufficient to permit their study at the university for a graduate degree through the medium of English. They had studied English for between 6 and 20 years, average 12.3. They had been immersed in English medium studies at the uni-

versity for between 1 month and 20 years, average 20.2 months. All participants were paid US$10 for taking part.

Results

For the native speakers, a forced entry multiple regression predicted final word VOT from word length, number of phonemes, frequency, and MI as the independent variables. It showed significant effects of number of phonemes (β = 0.31) and MI (β = –0.47—the greater the coherence of the formula, the more it primes access of its final word), but not of frequency. These data appear in Table 4.

The same analysis for the advanced ESL learners failed to show any significant predictors, although here too there was a relatively substantial effect of MI at β = –0.20.

Discussion

The continued evidence of an effect of formula MI on native speaker processing again strengthens its psycholinguistic validity. The more a formula coheres at greater than chance levels in the input, the more its last word is predicted by what comes before, and the more native speakers' language processing systems exploit these regularities in fluent processing. Advanced ESL learners also tend to reflect this result.

GENERAL DISCUSSION

The Psycholinguistic Validity of Corpus-Derived Formulas

Our results show that formulaic sequences, statistically defined and extracted from large corpora of usage, have clear educational and psy-

TABLE 4
Multiple Regressions Predicting Articulation Latency (Voice Onset Time) of the Formula's Final Word Following Priming in Native Speakers and Advanced ESL Learners in Experiment 3

Dependent variable Articulation latency (Voice onset time)	Predictors				
	n-gram length	n.phonemes	Frequency	Mutual information	R^2
Native English speakers					
β	0.11	0.31	–0.11	–0.47	11%
p	n.s.	0.03*	n.s.	0.01**	
Advanced ESL learners					
β	0.07	0.03	–0.08	–0.20	3%
p	n.s.	n.s.	n.s.	n.s.	

Note. *$p < 0.03$, **$p < 0.01$.

cholinguistic validity. 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.

Native speakers' language processing is affected by the MI of formulaic expressions when they are reading them for recognition of correct form (Experiment 1), reading them to access its pronunciation (Experiment 2), and reading aloud the final word after having processed the rest of the expression (Experiment 3).

Advanced ESL learners' language processing is affected by the frequency of formulaic expressions when they are reading them for recognition of correct form (Experiment 1), when reading them to access pronunciation (VOT, Experiment 2), and, marginally, when executing that articulation (articulation time, Experiment 2).

In sum, across a number of experiments, consistent evidence shows that formulaic expressions can be identified statistically from corpora of usage, and that native speakers and advanced ESL learners have become sensitive from their usage histories to these expressions so that they process them preferentially. But native speakers and learners are sensitive to different determinants of fluency—learners to *n*-gram frequency, fluent natives to MI.

Implications for a Model of Acquisition

The acquisition of linguistic knowledge and its fluent use, like other skills, is affected by frequency of exposure and practice. Thus, in the case of vocabulary, learners encounter high-frequency items more often than low-frequency items and tend to know them better. Proficiency tests can thus be stratified by frequency band, with more advanced learners being capable of answering lower frequency items (Nation, 2001). This is a simple result of input sampling: Learners must encounter a construction before they can consolidate a representation of its form and forge relevant meaningful associations. Advanced learners have had longer time on task and have thus encountered more constructions and more examples of each.

Consolidation of a recognition unit for a construction is only the beginning. It has to be tuned and made appropriately accessible (Ellis, 2006). In the acquisition of lexical and morphosyntactic fluency, processing speed can be explained simply by reference to the power law of learning (Anderson, 1982; DeKeyser & Sokalski, 1996; Ellis, 2002a; Ellis & Schmidt, 1998; Newell, 1990; Speelman & Kirsner, 2005). The power law of learning is generally used to describe the nonlinear relationships between practice and performance in a wide range of cognitive skills:

The effects of practice are greatest at early stages of leaning but eventually reach asymptote. Therefore, the effects of 10 additional exposures is very clear if a learner has only experienced that construction five times before, less marked if the learner has experienced it 50 times before, less discernible still if he or she has experienced it 500 times before.

There are thus two components of frequency effects (Ellis, in press). The first concerns chance of encounter—learners are more likely to encounter high-frequency constructions than low-frequency ones (Tomasello & Stahl, 2004). The second relates to the effects of practice on the strengthening of synaptic connections in the nervous system, with the effects of early practice increments being much more marked than those of later ones.

Consider then the effects of frequency of formula on processing in our native and nonnative speakers. The three strata of formula frequency in this study ranged from a low of 10.9 to a high of 43.6 occurrences per million. A back-of-the-envelope calculation for the native speakers is that they have been exposed to academic English for over 10 years at a language input rate of 30,000 words per day and an output of 7,500 words per day. This sums to 109 million words of input and 27 million words of output. Their baseline experience of the formulas used in this study might thus be between 1,188 and 4,572, their outputs between 294 and 1,177. There is surely plenty of scope for disputing these rough-and-ready estimates, but even if they are an order of magnitude out, it is still clear that for native language speakers, the vast majority of the formulas sampled here have been experienced plenty of times. They are all at the level where the practice function is leveling out, and there is little scope for discriminating between them.

For the nonnative speakers, though, most of whom are studying through the medium of English for the first time, things are quite different. Our participants had been immersed in English studies for perhaps 12 months or so on average, with the low end of the range in the three experiments being 1, 1, and 3 months. Their reading rates are slower than those of natives and the amount of input that becomes intake is far less than 100%. A generous estimate of intake of EAP materials of perhaps 10,000 words per day, sums over this period to 3.7 million words on average, but 300,000 words for the lowest exposure learners. Clearly, many of the learners will have experienced the formulas relatively infrequently, with the result that that they are still very much in the initial stages of tuning, where frequency effects are clearly discernable. Some of the learners may not have experienced some of these formulas at all.

We believe that these are the reasons why processing in learners is sensitive to frequency effects, while that in native speakers is not. What of the effects of MI?

Some recurrent multiword expressions exist simply by dint of being constituted of high-frequency words—examples from our corpus-derived lists include *and at the*, *that to the*, and even *the the um*. These word sequences appear in a wide variety of larger contexts, they are often grammatical fragments, and they do not have clearly discernable functions. They are high frequency but low MI. High-MI formulas, on the other hand, have much more clearly defined functions. Formulas very high on this measure have quite distinctive meanings, as technical phrases (e.g., *the citric acid cycle*, *nozzle melt pressure*, *the University of Michigan*), idioms (e.g., *on the one hand*, *come into play*, *ball-park estimate*), or constructions with clear discourse functions (e.g., the causative *that leads to*, the evaluative *it is interesting*, the contrastive *as opposed to*, the organizational *the first thing that*). They often constitute well-formed grammatical phrases. Their distinctive functions come from the recognition of them as coherent wholes. Higher MI *n*-grams are those with much greater coherence than is expected by chance. For the user to access their distinctive meanings, they need to be recognized as wholes, rather than interpreted openly and literally. The *citric acid cycle* has no everyday relation to the citric acid in our kitchen, *ball-park estimates* aren't restricted to ball parks, the cause-result *leads* in academic *that leads to* is very different from the *leads* typical of fiction whose subject and object are almost always animate. Our results show that native speakers are attuned to these constructions as packaged wholes. Their processing is a psycholinguistic instantiation of the idiom principle in that they preferentially recognize high-MI formulas as units.

Tuning the system according to frequency of occurrence alone is not enough for nativelike accuracy and efficiency. What is additionally required is tuning the system for coherence–for co-occurrence greater than chance. This is what solves the two puzzles for linguistic theory posed by Pawley and Syder (1983), nativelike selection and nativelike fluency. Native speakers have extracted the underlying co-occurrence information, often implicitly from usage; nonnatives, even advanced ESL learners with more than 10 years of English instruction, still have a long way to go in their sampling of language (Ellis, in press). They are starting to recognize and become attuned to more frequent word sequences, but they need help to recognize the distinctive formulas that are special to EAP.

APPLICATIONS TO LANGUAGE TEACHING AND LEARNING

Our ESL learners clearly need support in learning the formulaic sequences which have a high utility in the specialist discourse of EAP. Although these students have had more than a decade of English lan-

guage instruction, and, on average, perhaps 12 months of immersion in English-medium graduate education, they are still nonnativelike in their processing of formulas.

For the parallel case in EAP vocabulary, there have been long-standing 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 is simple: These items have the highest utility and should therefore be taught first. Corpus linguistic analyses of target texts and registers serve to identify what these learners most need to know. The language of general academic studies requires an AWL, a specialist vocabulary of 570 word families which increases coverage from the 78% provided by the 2,000 most frequent words of the language to a level of 87% sufficient for understanding general academic argument (Cobb, 2007; Coxhead & Nation, 2001; Nation, 1990, 2001).

The research described in this article shows that a supportive curriculum for ESL and EAP instruction should similarly identify an AFL and prioritize which formulas to teach. Other parts of our work (Simpson-Vlach and Ellis, in preparation) have pursued this goal. The AFL includes formulaic sequences, identifiable as frequently recurrent patterns in corpora of written and spoken discourse, that are significantly more common in academic discourse than in nonacademic discourse and which occupy a range of academic genres or habitats. It separately lists formulas that occur in both academic spoken and academic written language, as well as those that are more common in academic written language or in academic spoken language.

A major innovation that this research brings to the arena is a ranking of the formulas within the lists according to an empirically derived valid measure of utility, called *formula teaching worth* (FTW), which weighs formula frequency and MI in the same way as do skilled EAP instructors when judging usefulness for teaching. Most important, the AFL presents a classification of these formulas by pragmalinguistic function and offers some suggestions for including them in EAP curricula.

THE AUTHORS

Nick Ellis is a research scientist and professor of psychology at the University of Michigan, Ann Arbor, United States. His research interests include language acquisition, cognition, reading across languages, corpus linguistics, cognitive linguistics, and applied psycholinguistics. His current research focuses on second language acquisition, particularly usage-based acquisition and the probabilistic tuning of the system and emergentist accounts.

Rita Simpson-Vlach is a lecturer in the Department of Linguistics and Language Development at San José State University, California, United States. She was project

director of the Michigan Corpus of Academic Spoken English (MICASE) from its inception until 2006. Her research interests lie mainly in the area of corpus linguistics and EAP.

Carson Maynard joined the English Language Institute of the University of Michigan in 2003, where he is a lecturer in pronunciation, speaking, writing, and reading courses, and a program administrator. He spent a year as a researcher for the MICASE project.

REFERENCES

American Council on the Teaching of Foreign Languages (ACTFL). (1999). *ACTFL Proficiency Levels Revised 1999.* Alexandria, VA: ACTFL.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89,* 369–406.

Barlow, M. (2004). Collocate (Version 1.0) [Computer software]. Available from http://athel.com/product_info.php?products_id=29

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

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

BNC Consortium. (2001). BNC World [Database]. Available from http://www.natcorp.ox.ac.uk/.

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

Boersma, P., & Weenink, D. (2007). Praat (Version 4.6.01) [Computer software]. Available from http://www.fon.hum.uva.nl/praat/.

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

Bresnan, J. (1999, August). *Linguistic theory at the turn of the century. Plenary presentation.* Paper presented at the 12th World Congress of Applied Linguistics, Tokyo, Japan.

Bybee, J. (2003). Sequentiality as the basis of constituent structure. In T. Givón & B. F. Malle (Eds.), *The evolution of language out of pre-language* (pp. 109–132). Amsterdam: Benjamins.

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

Chafe, W. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing.* Chicago: University of Chicago Press.

Cobb, T. (2006). The Compleat Lexical Tutor (Version 4.5) [Computer software]. Available from http://132.208.224.131/.

Conklin, K., & Schmitt, N. (2007). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics, 28,* 1–18.

Cowie, A. P. (Ed.). (2001). *Phraseology: Theory, analysis, and applications.* Oxford: Oxford University Press.

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

Coxhead, A., & Nation, P. (2001). The specialised vocabulary of English for academic purposes. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for Academic purposes* (pp. 252–267). Cambridge: Cambridge University Press.

DeKeyser, R., & Sokalski, K. (1996). The differential role of comprehension and production practice. *Language Learning, 46,* 613–642.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition, 18,* 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*, 143–188.

Ellis, N. C. (2002b). Reflections on frequency effects in language processing. *Studies in Second Language Acquisition, 24*, 297–339.

Ellis, N. C. (2003). Constructions, chunking, and connectionism: The emergence of second language structure. In C. Doughty & M. H. Long (Eds.), *Handbook of second language acquisition* (pp. 63–103). Oxford: Blackwell.

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics, 27*, 1–24.

Ellis, N. C. (2008a). Phraseology: The periphery and the heart of language. In F. Meunier & S. Grainger (Eds.), *Phraseology in language learning and teaching* (pp. 1–13). Amsterdam: Benjamins.

Ellis, N. C. (2008b). The dynamics of second language emergence: Cycles of language use, language change, and first and second language acquisition. *Modern Language Journal, 92*, 232–249.

Ellis, N. C. (2008c). Usage-based and form-focused SLA: The implicit and explicit learning of constructions. In A. Tyler, K. Yiyoung, & M. Takada (Eds.), *Language in the context of use: Cognitive and discourse approaches to language* (pp. 93–120). Amsterdam: Mouton de Gruyter.

Ellis, N. C. (in press). Optimizing the input: Frequency and sampling in usage-based and form-focussed learning. In M. H. Long & C. Doughty (Eds.), *Handbook of second and foreign language teaching*. Oxford: Blackwell.

Ellis, N. C., Frey, E., & Jalkanen, I. (in press). The psycholinguistic reality of collocation and semantic prosody (1). Lexical access. In U. Römer & R. Schulze (Eds.), *Exploring the lexis-grammar interface*. Amsterdam: Benjamins.

Ellis, N. C., & Schmidt, R. (1998). Rules or associations in the acquisition of morphology? The frequency by regularity interaction in human and PDP learning of morphosyntax. *Language and Cognitive Processes, 13*, 307–336.

English Language Institute of the University of Michigan. (n.d.). The Michigan Corpus of Academic Spoken English. [Electronic version]. Retrieved April 2, 2006, from http://www.hti.umich.edu/m/micase

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text, 20,* 29–62.

Flowerdew, J., & Peacock, M. (Eds.). (2001). *Research perspectives on English for academic purposes*. Cambridge: Cambridge University Press.

Grabe, W., & Stoller, F. L. (2002). *Teaching and researching reading*. Harlow, England: Pearson Education.

Granger, S. (Ed.). (1998). *Learner English on computer*. London: Longman.

Granger, S., & Meunier, F. (Eds.). (in press). *Phraseology: An interdisciplinary perspective*. Amsterdam: Benjamins.

Gries, S. T., & Wulff, S. (2005). Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics, 3,* 182–200.

Hebb, D. O. (1949). *The organization of behaviour*. New York: John Wiley & Sons.

Hoey, M. P. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.

Hopper, P. J. (1998). Emergent grammar. In M. Tomasello (Ed.), *The new psychology of language: Cognitive and functional approaches to language structure* (pp. 155–176). Mahwah, NJ: Erlbaum.

Howarth, P. (1998). The phraseology of learners' academic writing. In A. P. Cowie

(Ed.), *Phraseology: Theory, analysis, and applications* (pp. 161–188). Oxford: Oxford University Press.
Hunston, S., & Francis, G. (1996). *Pattern grammar: A corpus driven approach to the lexical grammar of English.* Amsterdam: Benjamins.
Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing* (2nd ed.). Ann Arbor: University of Michigan Press.
ICAME. (1999). The ICAME [International Computer Archive of Modern and Medieval English] corpus collection (Version 2) [CD-ROM]. Norway: University of Bergen. Available from http://icame.uib.no/.
Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 229–254). Amsterdam: Benjamins.
Kuiper, K. (1996). *Smooth talkers: The linguistic performance of auctioneers and sportscasters.* Mahwah, NJ: Erlbaum.
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, England: Language Teaching.
Manning, C. D., & Schuetze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.
McDonald, S. A., & Shillcock, R. C. (2004). Eye-movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science, 14,* 648–652.
McLaughlin, B. (1995). *Fostering second language development in young children: Principles and practices* (Report No. 14). Santa Cruz: University of California, National Center for Research on Cultural Diversity and Second Language Learning. Retrieved June 10, 2008, from http://www.ncela.gwu.edu/pubs/ncrcdsll/epr14.htm
Meunier, F., & Granger, S. (Eds.). (2008). *Phraseology in language learning and teaching.* Amsterdam: Benjamins.
Nation, P. (1990). *Teaching and learning vocabulary.* Boston: Newbury House.
Nation, P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.
Nattinger, J. R. (1980). A lexical phrase grammar for ESL. *TESOL Quarterly, 14,* 337–344.
Nattinger, J. R., & DeCarrico, J. (1992). *Lexical phrases and language teaching.* Oxford: Oxford University Press.
Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics, 24,* 223–242.
Newell, A. (1990). *Unified theories of cognition.* Cambridge, MA: Harvard University Press.
Oakes, M. (1998). *Statistics for corpus linguistics.* Edinburgh, Scotland: Edinburgh University Press.
Ooi, V. B. Y. (1998). *Computer corpus lexicography.* Edinburgh, Scotland: Edinburgh University Press.
Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–225). London: Longman.
Psychology Software Tools. (2002). E-Prime (Version 1.1) [Computer software]. Available from http://www.pstnet.com/products/E-Prime/default.htm
Robinson, P., & Ellis, N. C. (Eds.). (2008). *A handbook of cognitive linguistics and second language acquisition.* London: Routledge.
Römer, U. (in press). Corpora and language teaching. In A. Lüdeling & M. Kytö

(Eds.), *Korpuslinguistik—Corpus linguistics. An international handbook*. Berlin: Mouton de Gruyter.
Schmitt, N. (Ed.). (2004). *Formulaic sequences*. Amsterdam: Benjamins.
Simpson, R., & Mendis, D. (2003). A corpus-based study of idioms in academic speech. *TESOL Quarterly, 3,* 419–441.
Simpson-Vlach, R., & Ellis, N. C. (in press). An academic formulas list (AFL).
Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Sinclair, J. (Ed.). (1996). *How to use corpora in language teaching*. Amsterdam: Benjamins.
Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.
Speelman, C., & Kirsner, K. (2005). *Beyond the learning curve: The construction of mind*. Oxford: Oxford University Press.
Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
Swinney, D. A., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior, 18,* 523–534.
Tomasello, M., & Stahl, D. (2004). Sampling children's spontaneous speech: How much is enough? *Journal of Child Language, 31,* 101–121.
University of Pennsylvania Linguistic Data Consortium. (n.d.). Switchboard [English corpus]. Philadelphia: University of Pennsylvania, Author. Available from http://www.ldc.upenn.edu/. Switchboard User's Manual is available from http://www.ldc.upenn.edu/Catalog/readme_files/switchboard.readme.html.
West, M. (1953). *A general service list of English words*. London: Longman.
Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.