The development of formulaic sequences in first and second language writing

Investigating effects of frequency, association, and native norm*

Matthew Brook O'Donnell[1], Ute Römer[2] and Nick C. Ellis[1]
[1]University of Michigan / [2]Georgia State University

Formulaic sequences are recognised as having important roles in language acquisition, processing, fluency, idiomaticity, and instruction. But there is little agreement over their definition and measurement, or on methods of corpus comparison. We argue that replicable research must be grounded upon operational definitions in statistical terms. We adopt an experimental design and apply four different corpus-analytic measures, variously based upon n-gram frequency (Frequency-grams), association (MI-grams), phrase-frames (P-frames), and native norm (items in the Academic Formulas List – AFL-grams), to samples of first and second language writing in order to examine and compare knowledge of formulas in first and second language acquisition as a function of proficiency and language background. We find that these different operationalizations produce different patterns of effect of expertise and L1/L2 status. We consider the implications for corpus design and methods of analysis.

Keywords: formulaic sequences, n-grams, phrase-frames, academic writing, proficiency development

1. Introduction

1.1 Formulaic sequences

Corpus linguistics demonstrates that much of communication makes use of formulaic sequences, that language is rich in collocational and colligation restrictions and semantic prosodies, and that the phrase is the basic level of language

International Journal of Corpus Linguistics 18:1 (2013), **83–108.** DOI 10.1075/ijcl.18.1.070do
ISSN 1384–6655 / E-ISSN 1569–9811

representation where form and meaning meet with greatest reliability (Römer 2009a; Sinclair 1991, 2004, 2005; Stubbs 2001, 2007).

Formulaic sequences are also the focus of a wide range of research across applied linguistics (Cowie 2001, Ellis 2008, Granger & Meunier 2008, O'Donnell & Römer, in preparation, Römer 2010, Schmitt 2004, Wray 2002), cognitive linguistics (Gries & Wulff 2005, Robinson & Ellis 2008) and psycholinguistics (Ellis 2012). Formulaic sequences are considered central in idiomatic and fluent *use of language* (Ellis 1996, Erman & Warren 2000, Nattinger & DeCarrico 1992, Pawley & Syder 1983), as well as in *language acquisition,* with learners progressing from the use of memorised, high functional-utility formulaic phrases, through slot-and-frame patterns, to more open grammatical constructions (Ellis 2003, Ellis & Cadierno 2009, Tomasello 2003). Thus, formulaic sequences are thought to be psycholinguistically real. In the words of two leading corpus linguists: recurrent chains are "one type of evidence of abstract cognitive schemas" (Stubbs & Barth 2003: 84); and "[a]ll lexical items are primed for grammatical and collocational use, i.e. every time we encounter a lexical item it becomes loaded with the cumulative effects of these encounters, such that it is part of our knowledge of the word that it regularly co-occurs with particular other words or with particular grammatical functions" (Hoey 2004: 21; 2005).

1.2 Corpus analysis of formulaic sequences

Every genre of *English for Academic Purposes* (EAP) and *English for Specific Purposes* (ESP) has its own phraseology, and learning to be effective in the genre involves learning the relevant formulas (Swales 1990). EAP research (e.g. Biber et al. 2004, Flowerdew & Peacock 2001, Hyland 2004, Swales 1990) therefore focuses on determining the functional patterns and constructions of different academic genres. Stubbs & Barth (2003: 84) described how corpus studies may help us understand "how the behaviour of many different speakers can become co-ordinated and focussed around the norms that we recognize as the idiomatic language use of a speech community". One example of such corpus research is that of Simpson-Vlach & Ellis (2010) who analysed the spoken and written language of the university 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).[1] The AFL includes formulaic sequences identified as (i) frequent recurrent patterns in corpora of written and spoken language, which (ii) occur significantly more often in academic than in non-academic discourse, and (iii) inhabit a wide range of academic genres. It separately lists formulas that are common in academic spoken

and academic written language (for example *we can see, to some extent, an example of*), as well as those that are special to academic written language alone (e.g. *on the other hand, as can be seen, be argued that*), and academic spoken language alone (e.g. *does that make sense, you might want to, you know what I mean*). Ellis & Simpson-Vlach (2009) assess the educational validity of the AFL items, showing that experienced instructors rated formulas as being bona fide phrases worthy of teaching if they were (1) strongly internally associated, (2) of high frequency, and (3) had a cohesive meaning or function as a phrase. Simpson-Vlach & Ellis (2010) classified the formulas according to their predominant pragmatic function for descriptive analysis and in order to marshal the AFL for inclusion in English for Academic Purposes instruction.

1.3 Psycholinguistic analysis of formulaic sequences

Psycholinguistic research independently investigates the psychological reality of formulaic sequences using experimental methods. Swinney & Cutler (1979) found that subjects took much less time to judge idiomatic expressions, such as *kick the bucket*, as being meaningful English phrases than they did for non-idiomatic control strings like *lift the bucket* (see also Conklin & Schmitt 2008, Schmitt 2004). Bod (2001), using a lexical-decision task, showed that high-frequency three-word sentences such as *I like it* were reacted to faster than low-frequency sentences such as *I keep it* by native speakers. Durrant & Doherty (2010) used lexical decision to demonstrate that the first word of low- (e.g. *famous saying*), middle- (*recent figures*), high-frequency (*foreign debt*) and high-frequency and psychologically-associated (*estate agent*) collocations primed the processing of the second. Arnon & Snider (2010) used a phrasal decision task ("is this phrase possible in English or not?") to show that comprehenders are also sensitive to the frequencies of compositional four-word phrases: more frequent phrases (e.g. *don't have to worry*) were processed faster than less-frequent phrases (*don't have to wait*) even though these were matched for the frequency of the individual words or substrings. Tremblay et al. (2012) examined the extent to which lexical bundles (LBs, defined as frequently recurring strings of words that often span traditional syntactic boundaries) are stored and processed holistically. Three self-paced reading experiments compared sentences containing LBs (e.g. *in the middle of the*) and matched control sentence fragments (*in the front of the*) such as *I sat in the middle/front of the bullet train*. LBs and sentences containing LBs were read faster than the control sentence fragments in all three experiments.

Maintenance of material in short-term memory and accurate subsequent production is also affected by knowledge of formulaic sequences. As Bybee (2005)

quips (after Hebb's (1949) "Cells that fire together, wire together") "Words used together fuse together." Jurafsky et al. (2001) analyzed the articulation time of successive two-word sequences in the Switchboard corpus to show that in production, humans shorten words that have a higher contextualized probability. Bannard & Matthews (2008) identified frequently occurring chunks in child-directed speech (e.g. *sit in your chair*) and matched them to infrequent sequences (e.g. *sit in your truck*). They tested young children's ability to produce these sequences in a sentence-repetition test. Three-year-olds and two-year-olds were significantly more likely to repeat frequent sequences correctly than infrequent sequences. People have longer-term memory as well for the particular wording used to express something (as any parent who misreads a favourite bed-time story can readily attest). Some learning takes place after just one incidental exposure. Gurevich et al. (2010) showed that adult native speakers recognize at above chance rates full sentences that they have been exposed to only once in texts of 300 words long that were presented non-interactively with no advanced warning of a memory test. Even after a six-day delay, participants reliably reproduced sentences they have heard before when asked to describe scenes, even though they were not asked to recall what they had heard.

1.4 Triangulating corpus and psycholinguistic research

There have been some attempts to triangulate *corpus* and *psycholinguistic* approaches to formulaic language. Ellis & Simpson-Vlach (2009) and Ellis et al. (2008) used four experimental procedures to determine how the corpus linguistic metrics of frequency and mutual information (MI, a measure of the degree to which the words of a formula attract each other) are represented implicitly in native and non-native speakers, thus to affect their accuracy and fluency of processing of the formulas of the AFL. The language processing tasks in these experiments were selected to sample an ecologically valid range of language processing skills: spoken and written, production and comprehension, form-focused and meaning-focused. They were: (i) speed of reading and acceptance in a grammaticality judgment task where half of the items were real phrases in English and half were not, (ii) rate of reading and rate of spoken articulation, (iii) binding and primed pronunciation – the degree to which reading the beginning of the formula primed recognition of its final word, (iv) speed of comprehension and acceptance of the formula as being appropriate in a meaningful context. Processing in all experiments was affected by various corpus-derived metrics: length, frequency, and MI. Frequency was the major determinant for non-native speakers, but for native speakers it was predominantly the MI of the formula which determined processability.

1.5 Operationalizing formulaic language

Despite this widespread interest in formulaic language, there is little agreement about how it might be operationally defined and measured. A widely cited general definition of a formulaic sequence is that of Wray (2000: 465): "A sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is stored and retrieved whole from the memory at the time of use, rather than being subject to generation or analysis by the language grammar". This definition, while admirably encompassing a wide range of research interests as indeed it was designed to do, is too open to allow tightly replicable research. Equally, within corpus linguistics, "there is no purely automatic way of identifying phrasal units of meaning" (Stubbs 2007: 181).

We believe that formulaic language should be more objectively defined in terms of measureable operationalizations. To this end, our research here explores corpus-linguistic techniques that provide a range of methods for the quantification of recurring sequences (e.g. clusters, n-grams, collocations, phrase-frames) and gauging the strength of association between the component words.[2]

Corpus research often results in equivocal findings too because it is hard to know whether a difference in point estimates between two corpora is important or not. Normalised frequency counts on just about any feature are going to differ, so when are differences significant or not? This is a recurrent question on the Corpora List (corpora@uib.no; see http://www.hit.uib.no/corpora/). Therefore our research also adopts experimental methods, taking multiple independent random samples from the corpora of interest. This allows statistical inference, comparing between- and within-corpus variance, to test whether any between-corpus differences are statistically significant.

1.6 The goals of this research

The larger research goal is one of convergent and differential validation of operationalization. Different methods for identifying formulaic sequences need to consider both contiguous sequences (n-grams, e.g. *at the end of, on the other hand*), limited span frameworks (phrase-frames, e.g. *at the * of, on the * hand* and skip-grams, e.g. *at end of*, *on other hand*, *on hand*; Fletcher 2002–2007, Stubbs 2007, Guthrie et al. 2006) and collocational clusters (item-sets and concgrams; Cheng et al. 2006). Relevant statistical measures include frequency above a certain threshold, MI, t-score, gravity counts (Daudaravičius & Marcinkevičienė 2004), and various reference lists of formulas and various dispersion measures.

We want to determine how the different measures reflect a range of independent variables, including text- and corpus size, number of authors, register (spoken/written) and genre differences (e.g. news, fiction, academic prose), vocabulary diversity (such as type-token entropy), speaker age, nativeness and proficiency. We believe these are the steps necessary to usefully further the standardization of the measurement of formulaicity in language usage and that they will prove the foundations for triangulation with psycholinguistic definitions of formulaicity as well as for studies of first and second language acquisition, instruction, and evaluation.

Here we report just one part of this larger project, one that focuses upon formulaic language in academic writing. The current study develops a variety of statistical measures of formulaic language and applies them in the study of the development of L1 and L2 academic writing. Corpora of writing are sampled from different L1 backgrounds and at a range of proficiency levels, including: European University English learner writing (ICLE; International Corpus of Learner English, Granger 1998), undergraduate native English student writing (LOCNESS; Louvain Corpus of Native English Essays, Granger 1998), A-graded graduate writing by non-native English speakers (MICUSP-NNS; Michigan Corpus of Upper-level Student Papers, Römer & O'Donnell 2011), A-graded final year undergraduate and graduate writing by native English speakers (MICUSP-NS; Römer & O'Donnell 2011), and published research articles written by native or near-native English speakers (Hyland 1998, 2004).[3] We take eight independent random samples from each of these corpora. Using these, we quantify and compare the frequencies and learner uptake of continuous sequences of various lengths (n-grams, e.g. *at the end of*) and associated 'frames' (e.g. *at the * of*). Various statistical analyses are applied to investigate the effects of (a) proficiency development in the usage of these units, and (b) L1 backgrounds.

Our previous psycholinguistic work focused upon implicit knowledge of formulaic language as evidenced in on-line processing tasks (Ellis & Simpson-Vlach 2009, Ellis et al. 2008). Here we investigate written language production which allows authors more scope for the conscious construction of language, and the use of explicit knowledge to create, consider, and reshape their argument and its means of expression. Nevertheless, we expect to find effects of (1) expertise, whereby more proficient writers show more use of formulas, and (2) native language status. We also expect that different statistical operationalizations of formulaic language will produce different patterns of results.

2. Measuring formulaic language in apprentice and expert academic writing

2.1 Measures

Four immediate options for the basis of determination of formulaic sequences come to mind: n-gram frequency, n-gram association, phrase-frames, or native norms. Let us consider each in turn.

2.1.1 *Frequency*

Formulas are recurrent sequences. One definition, then, is that we should identify strings that recur often. This is the *lexical bundle* approach of Biber and colleagues (Biber 2006, Biber et al. 2004, Biber et al. 1999), based solely on frequency. It has the great advantages of being methodologically straightforward and having face validity – we all readily agree that high-frequency strings like *how are you*, *nice day today*, and *good to see you* are formulaic sequences. There is psycholinguistic support for such measures too – the last 50 years of psycholinguistic research have demonstrated language processing to be exquisitely sensitive to usage frequency at all levels of language representation: phonology and phonotactics, reading, spelling, lexis, morphosyntax, formulaic language, language comprehension, grammaticality, sentence production, and syntax (Ellis 2002). Language knowledge involves statistical knowledge, so humans learn more easily and process more fluently high frequency forms and "regular" patterns which are exemplified by many types and which have few competitors. But we also know some formulas that are not high frequency, like *blue moon*, *longitude and latitude*, and *raining cats and dogs*. And other high-frequency strings, like *and of the*, or *but it is*, don't seem very formulaic. Definitions in terms of frequency alone result in long lists of recurrent word sequences that collapse distinctions that intuition would deem relevant. 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. The fact that a sequence of words is above a certain frequency threshold does not necessarily imply either psycholinguistic salience or pedagogical relevance (Ellis et al. 2009, Leech 2011).

2.1.2 *Association*

Psycholinguistically salient sequences, on the other hand, like *on the other hand*, *longitude or latitude*, or *raining cats and dogs*, 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. There are numerous statistical

measures of association available, each with their own advantages and disadvantages (Evert 2005, Gries 2013, Gries & Divjak 2012, Gries & Stefanowitsch 2004, Stubbs 1995). In this paper, as in our previous work with the AFL, we use Mutual Information (MI). 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 & Schütze 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. MI privileges coherent strings that are constituted by low frequency items (Evert 2005).

2.1.3 *Phrase-frames*

Our first two measures define formulas as frozen n-gram sequences, but some sequences can be much more similar than others. Consider semi-productive patterns such as *it is * to* where the framing elements surround a variable slot that can be filled by a range of different words but not just *any* word. In the case of *it is * to*, words that commonly fill the gap include *interesting*, *good*, *useful*, and *nice*. A term commonly used to refer to such sequences of words which contain a variable slot is 'phrase-frame', short 'p-frame' (e.g. Römer 2010, Stubbs 2007). P-frames provide systematic groupings of n-grams which are identical except for one word in the same position (e.g. *it is interesting to*, *it is good to*, *it is useful to* and *it is nice to* are summarized under the p-frame *it is * to*). The words that fill the * slot are the 'variants' of the p-frame. Looking at p-frames and the type and token frequencies of their variants can provide insights into the variability of formulaic sequences. It helps us see to what extent Sinclair's Idiom Principle (Sinclair 1987, 1991, 1996) is at work and how fixed language units are or how much they allow for variation. Note also that p-frame variants are often semantically related, as in the case of *interesting*, *good*, *useful*, and *nice*. Examples of two common p-frames and their most frequent variants in one of our datasets (from MICUSP_NS) are given in Table 1.

2.1.4 *Native norms*

Definitions purely in terms of frequency or association might well reflect that the language production makes use of sequences that are ready-made by the speaker or writer, but these may not necessarily be nativelike. Non-native academic writing can often be identified by the high frequency of use of phrases that come from strategies of translation (like *make my homework*, or *make a diet*), or formulas that

Table 1. Examples of p-frames with common variants in MICUSP_NS

the * of the
the end of the
the rest of the
the beginning of the
the nature of the
the results of the
it is * that
it is clear that
it is possible that
it is likely that
it is important that
it is true that

occur frequently in spoken language but which are frowned upon as informal in academic writing (like *have a nice day!*, or *it is stupid to…*). An additional, divergent, criterion for formulas is that they reflect *native-like* selection and *native-like* fluency (Pawley & Syder 1983), and we measure academic writing by how well it uses the formulaic sequences and grammatico-lexical techniques of the norms of its reference genre. Our operationalization of this criterion makes use of the AFL from our prior research (Ellis & Simpson-Vlach 2009, Ellis et al. 2008, Simpson-Vlach & Ellis 2010). We search for the instances of these academic patterns in our corpora of native and non-native English academic writing at different levels of proficiency.

2.2 Corpora used

As described above, we draw upon a range of corpora that capture academic writing by native and non-native speakers at different academic levels. These corpora, listed in Table 2, cut across the factors of nativeness and proficiency (or expertise) and allow insights into writing development. We distinguish between collections of apprentice or novice academic texts and expert academic texts. Following Scott & Tribble (2006: 133), we understand apprentice texts to be "unpublished pieces of writing that have been written in educational or training settings", whereas expert texts are pieces of writing that have been published. We assume that the level of expertise corresponds with the level of academic writing proficiency.

At the lowest level of proficiency of the selected corpora, the eleven components of ICLE (first release) capture non-native, predominantly undergraduate student writing by learners of different first language backgrounds (Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian,

Spanish, Swedish).[4] Each subcomponent of ICLE consists of between 260 and 460 short student argumentative essays. Despite somewhat varying numbers of texts, the eleven parts of ICLE are of comparable size, containing between 200,000 and 287,000 word tokens (for detailed figures, see Table 2). LOCNESS serves as a comparative corpus for ICLE. It contains similar types of texts as the learner corpus, only they were written by native American and British university students and British A-level pupils. In order to ensure comparability with ICLE, we excluded all A-level texts from our analysis and only used a LOCNESS subset of 365 undergraduate student essays, making up around 269,000 words of running text. Matching LOCNESS and the ICLE subcorpora in size but capturing writing at a more advanced academic level by more proficient student writers, MICUSP-NS and MICUSP-NNS present another pair of datasets that enables comparisons of native and non-native speaker production. The NS and NNS texts selected from MICUSP for these two datasets are all successful (A-graded) assignments submitted by graduate students from a range of academic disciplines: Biology, Civil and Environmental Engineering, Economics, Education, English, History and Classical Studies, Industrial and Operations Engineering, Linguistics, Mechanical Engineering, Natural Resources and Environment, Nursing, Philosophy, Physics, Political Science, Psychology and Sociology. The most common types of texts in MICUSP-NS and MICUSP-NNS are argumentative essay, report and research paper.

Table 2. Overview of corpora/datasets

Corpus		Texts	Tokens*
LOCNESS		365	269,839
ICLE	Bulgarian	302	200,905
	Czech	258	207,739
	Dutch	288	271,411
	Finnish	391	275,944
	French	460	287,699
	German	450	234,620
	Italian	398	226,984
	Polish	365	235,065
	Russian	279	232,035
	Spanish	260	205,003
	Swedish	371	208,408
MICUSP-NNS		72	256,318
MICUSP-NS		62	256,314
HYLAND		61	256,916

* Token counts from WordSmith Tools Version 5.0.

All corpora described so far can be considered apprentice corpora. The students who contributed to ICLE, LOCNESS and MICUSP are novice academic writers on their way to developing expert writing skills and becoming accepted members of academic communities of practice. Our study compares these sets of NS and NNS apprentice academic writing at different stages with a collection of expert writing culled from the Hyland corpus of published research articles (Hyland 1998) and here referred to as HYLAND. All HYLAND authors are, in Swales' (Swales 2004: 56) terms, "senior" scholars, have had multiple-year university training, and succeeded in getting published in peer-reviewed journals. For the HYLAND subset, we selected 61 texts from the eight disciplinary subsections of the Hyland corpus: Cell Biology, Electrical Engineering, Linguistics, Marketing, Mechanical Engineering, Philosophy, Physics and Sociology. We selected texts from each of the eight disciplines in a round-robin fashion using the lowest word counts in an attempt to end up with as many texts as possible. The 61 selected texts make up about 256,916 words.[5]

2.3 Stratified sampling, n-gram/p-frame extraction and formula matching

2.3.1 *Corpus sampling approach*

In much corpus-based research of lexico-grammatical variation across text-types, genres and registers – what might be termed 'the variationist approach' – the central methodological concern has been assembling reasonably matched (in terms of size) corpora for each type of sublanguage in the comparison. For example, in a comparison of newspaper, academic, fiction and spoken language, the common approach would be to create four (sub)corpora of reasonable size, to count some feature and then normalize these counts (e.g. occurrences per million words). Clearly, when dealing with language, and particularly for lexical concerns, the appropriate level of granularity for such comparisons is not the single text but can be achieved by pooling together similar texts. This approach tends to result in just a single measure, of occurrences per million words for instance, for each of the comparison parts. However, less attention has been given to the possibility of internal variation within one of these corpora, for instance, a particular newspaper text or set of texts within the newspaper corpus that exhibit a much higher or lower frequency of a particular feature than the rest of the texts (cf. Gries 2006 on importance of including dispersion measures in reporting corpus frequencies). These outliers can greatly affect the arithmetic mean used to calculate occurrences per million words. The experimental design adopted for this study prompted us to carry out a sampling procedure to derive subsamples within each of our corpora of interest (representing learner level and (non-)nativeness). Using multiple samples for each group both helps identify potential outliers in terms of frequency of the

feature or features of interest that might skew a single average value and allows us to utilize statistical procedures, such as ANOVA, to identify significant differences in our formula measures between the various corpora.

The corpora selected for the study were discussed in Section 2.2 and the number of texts and token counts for each are shown in Table 2. The number of texts per corpus range from 61 (HYLAND) to 460 (ICLE French) (median 302) and from 200,905 (ICLE Bulgarian) to 287,699 (ICLE French) tokens (median 235,065). We determined that dividing each of our corpora into eight subcorpus samples would still result in text counts and token counts large enough to allow enough repetition to produce a sizeable number of repeated 3-, 4- and 5-grams. We considered a number of sampling strategies, such as dividing each corpus precisely into eighths according to token count – which would result in a variant number of texts and partial texts in each sample – or according to text count – which would result in samples with variant token counts. The desired strategy is one that would result in (as closely as possible) both the same number of texts and the same overall token count in each sample. By analogy, imagine taking a pack of 52 cards and being asked to deal it out into 8 piles, each with the same number of cards in it and with the summed values of each of the cards in the pile being equal.[6] On the face of it this would appear to be a relatively straightforward procedure that, as Hayes (2002) points out, is carried out by children in school playgrounds on a daily basis while picking teams/sides for a game. However, it falls into the class of algorithmic problems classified as NP-complete, i.e. computationally intractable! Fortunately, there are a number of heuristic algorithms that provide approximate solutions to this problem (formally known as the Number Partitioning Problem (NPP)). We adapted the algorithm developed by Karmarker & Karp (1983), which, on each iteration, divides a set of integers into two sets of equal size (for an even number of members) with the same or nearly the same sum totals. Table 3 shows the result of applying this algorithm to the 460 texts in ICLE French.[7]

Table 3. Eight samples for ICLE French corpus

Group	Number of files	Tokens*
1	56	36,154
2	58	36,154
3	58	36,154
4	56	36,155
5	59	36,154
6	59	36,154
7	57	36,155
8	57	36,154
Total	**460**	**289,234**

* Token counts calculated using a simple whitespace tokenizer.

Six of the eight samples have a total of 36,154 tokens and the other two just one token more with 36,155. This is remarkable given that none of the texts are split. The number of texts per sample range from 56 to 59. This procedure was repeated for each of the corpora in Table 2.

2.3.2 *N-gram extraction*

In the initial data extraction, we collected n-grams of length n = 2 up to n = 9. However, on the basis of initial ANOVA only values of n = 2, 3, 4 and 5 were shown to have significance. As p-frame analysis requires initial n-grams of n > 2, in this study we restricted the analysis to 3-, 4- and 5-grams. We used the Word List tool in WordSmith Tools Version 5.0 (Scott 2008) to build an index of each of the eight samples for each corpus and then compute clusters of length 3–5 with a minimum frequency of three. For each sample the number of n-gram types was recorded and this was normalized to types per million tokens according to the token count for each sample. This number was used as the *Frequency-grams* measure in the subsequent data analysis.

2.3.3 *Mutual Information calculation*

We chose to use MI as our n-gram association measure (see 2.1.2). Mutual Information is not a significance statistic and therefore does not have a table of threshold values for associated p-values. In order to use MI as one of our formulaic measures we needed to introduce threshold values for each value of n. Different thresholds are needed because MI-values increase in a non-linear manner according to n-gram length. We derived our thresholds by extracting n-grams (n = 3, 4 and 5) with a frequency 3+ from the one million word BNC Baby Academic corpus and calculating MI values for each n-gram. The median values for each n were then selected as our MI thresholds to use in subsequent analysis (3-grams: 6.723, 4-grams: 13.085 and 5-grams: 20.835). We calculated an MI-score for all the n-grams extracted for each sample for each corpus as described in 2.3.2 and discarded those with scores below the appropriate threshold. The number of types for the remaining n-grams was recorded and normalized to types-per million tokens. This number was used as the *MI-grams* measure in the subsequent data analysis.

2.3.4 *P-frame extraction*

The P-frame measure was calculated using the output of the KfNgram tool (Fletcher 2002–2007). For each of the eight samples for each of the corpora, all 3-, 4- and 5-grams were extracted with a frequency threshold of 1 (i.e. ALL possible 3–5 grams). In the second step, the frequency threshold was set to 3 and 3-, 4- and 5-p-frames were extracted from the n-gram lists produced in the first step (e.g. *the * of*, *the * of the*, *at the * of the*, where variants include {*end*, *beginning*}). For

each sample, the number of p-frames of each length of n was recorded and normalized to types per million tokens according to the token count for each sample. This number was used as the *P-frames* measure in the subsequent data analysis.

2.3.5 *AFL n-gram calculation*

Using the n-gram lists extracted as described in Section 2.3.2, we made a comparison with the Academic Formula List (AFL) written and core components (see 2.1.4).[8] The types in each n-gram list were intersected with those in the AFL_{core} + $AFL_{written}$ and the number of n-gram types was recorded. These type counts were then normalized to types per million tokens according to the token count for each sample. This number was used as the *AFL-grams* measure in the subsequent data analysis.

3. Results

3.1 Frequency-defined N-grams across corpora (Frequency-grams)

Figure 1 shows boxplots for each of the corpora of the type frequency of 3, 4 and 5 Frequency-grams occurring three or more times. A two factor ANOVA (15 Corpora × 3 Ngram-Length) demonstrated significant effects of Corpora (F[14,315] = 47.10, p < .001), Length (F[2,315] = 4635.45, p < .001), and a significant Corpus × Length interaction (F[22,315] = 5.17, p < .001). The Expert and Graduate writers look in this plot to be producing more frequency-defined n-grams than the LOCNESS and ICLE writers, and these differences are more pronounced with longer n-grams.

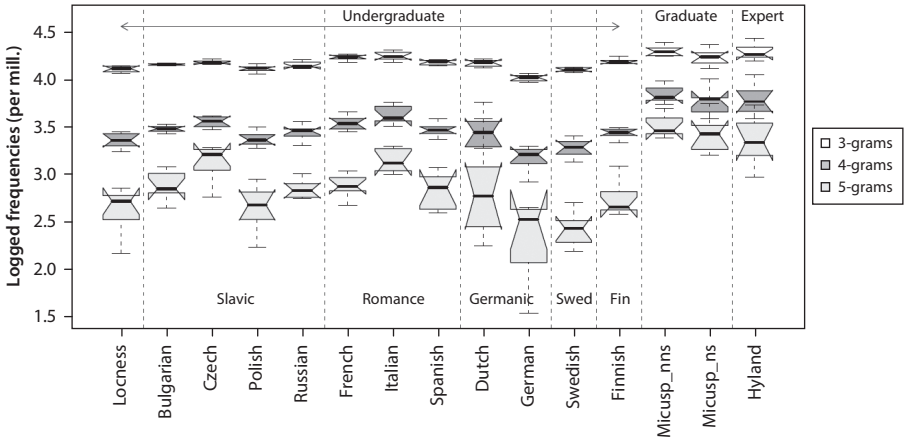

Figure 1. Type frequency of 3-, 4-, and 5-grams using 3+ frequency threshold (Frequency-grams)

We regrouped the ICLE L2 writers into one pool in order to assess this further. We show the model means in the upper left hand panel of Figure 5. A two factor ANOVA (5 Groups × 3 Ngram-Length) demonstrated significant effects of Group (F[4,345] = 73.46, p < .001), Length (F[2,345] = 2781.40, p < .001), and a significant Group × Length interaction (F[8,345] = 4.09, p < .001). Post hoc testing using Tukey's Honestly Significant Differences (HSD) test showed no significant differences between the Expert L1 writers (HYLAND) and the Graduate writers (MICUSP_NS, MICUSP_NNS). It also showed that all of these groups outperformed the native (LOCNESS) and L2 (ICLE) undergraduate writers, and that the L2 writers were producing marginally more Frequency-grams than their L1 peers (MICUSP_NS vs. MICUSP_NNS, p < .05; LOCNESS vs. ICLE L2, p = .08).

Thus, for frequency-defined formulas, there are clear effects of expertise (Expert ≈ Graduate > Undergraduate), with, if anything, L2 learners producing more formulas than their native speaker peers.

3.2 MI-defined N-grams across corpora (MI-grams)

Figure 2 shows boxplots for each of the corpora of the type frequency of 3, 4 and 5 MI-grams occurring with an MI > 6.723 (3-grams), 13.085 (4-grams) or 20.835 (5-grams) (see Section 2.3.3). A two factor ANOVA (15 Corpora × 3 Ngram-Length) demonstrated significant effects of Corpora (F[14,315] = 8.59, p < .001), Length (F[2,315] = 244.20, p < .001), and a significant Corpus × Length interaction (F[22,315] = 4.42, p < .001). The Expert and Graduate writers look in this plot to be producing more MI-grams than the LOCNESS and ICLE writers and these differences are more pronounced with longer n-grams.

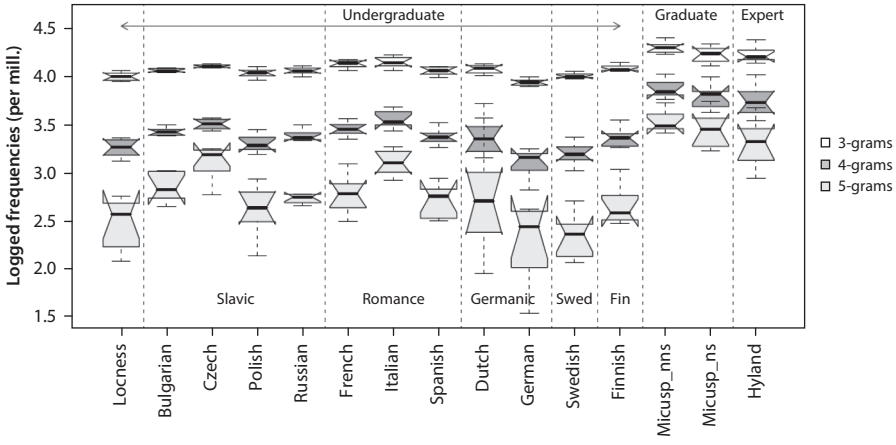

Figure 2. Type frequency of 3-, 4-, and 5-grams using association threshold (MI-grams)

A two factor ANOVA (5 Groups × 3 Ngram-Length) demonstrated significant effects of Group ($F[4,345] = 27.09$, $p < .001$), Length ($F[2,345] = 250.85$, $p < .001$), and a significant Group × Length interaction ($F[8,345] = 15.12$, $p < .001$). We show the model means in the upper right hand panel of Figure 5. Post hoc testing using Tukey's Honestly Significant Differences (HSD) test showed no significant differences between the Expert L1 writers (HYLAND) and the A grade University writers (MICUSP_NS, MICUSP_NNS). It also showed that all of these groups outperformed the native (LOCNESS) and L2 (ICLE) undergraduate writers, and that the L2 writers were producing roughly the same amount of MI-defined formulas as their L1 peers (MICUSP_NS vs. MICUSP_NNS, *ns*; LOCNESS vs. ICLE L2, *ns*). Thus, for MI-defined formulas, there are clear effects of expertise (Expert ≈ A grade graduate > Undergraduate), but there is no effect of L1/L2 status.

3.3 P-frames across corpora

Figure 3 shows boxplots for each of the corpora of the type frequency of 3, 4 and 5-p-frames. A two factor ANOVA (15 Corpora × 3 Ngram-Length) demonstrated no differences between the Corpora ($F[14,315] = 1.77$, *NS*), but significant effects of Length ($F[2,315] = 630.38$, $p < .001$), and a significant Corpus × Length interaction ($F[28,315] = 9.64$, $p < .001$). For the longer p-frames, the Expert and Graduate writers look in this plot to be producing roughly equivalent numbers of p-frames, and more than the LOCNESS and ICLE writers.

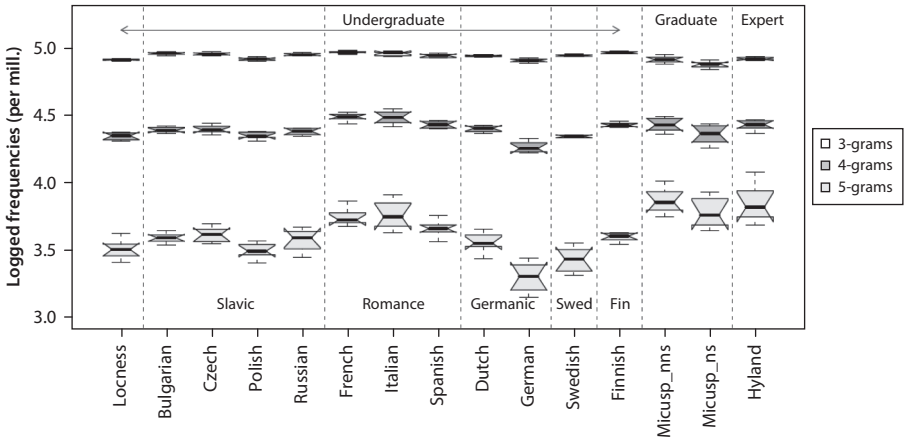

Figure 3. Type frequency of 3-, 4-, and 5-p-frames using 3+ frequency threshold (P-frames)

A two factor ANOVA (5 Groups × 3 Ngram-Length) demonstrated no effects of Group ($F[4,345] = 0.94$, *ns*), but significant effects of Length ($F[2,345] = 669.72$, $p < .001$), and a significant Group × Length interaction ($F[8,345] = 35.53$, $p < .001$). We show the model means in the bottom left hand panel of Figure 5. Post hoc testing using Tukey's Honestly Significant Differences (HSD) test showed no significant differences overall between any of the groups of writers. Thus, for P-frames, there are neither significant effects of expertise nor of L1/L2 status (cf. Römer 2009b, which makes use of p-frames to investigate proficiency in academic writing).

3.4 Target formulas across corpora (AFL-grams)

We pool the AFL items of different lengths together since there are few 5-grams. Figure 4 shows boxplots for each of the corpora of the type frequency of AFL-grams. A one-way ANOVA (15 Corpora) demonstrated significant differences between these ($F[14,105] = 12.09$, $p < .001$). The Expert writers look to be producing more AFL-grams than both the Graduate writers and the LOCNESS and ICLE writers, with little discernable difference between these latter corpora.

A one-way ANOVA (5 Groups) demonstrated significant effects of Group ($F[4,345] = 6.03$, $p < .001$). We show the model means in the lower right hand panel of Figure 5. Post hoc testing using Tukey's Honestly Significant Differences (HSD) test showed significant differences between the Expert L1 writers

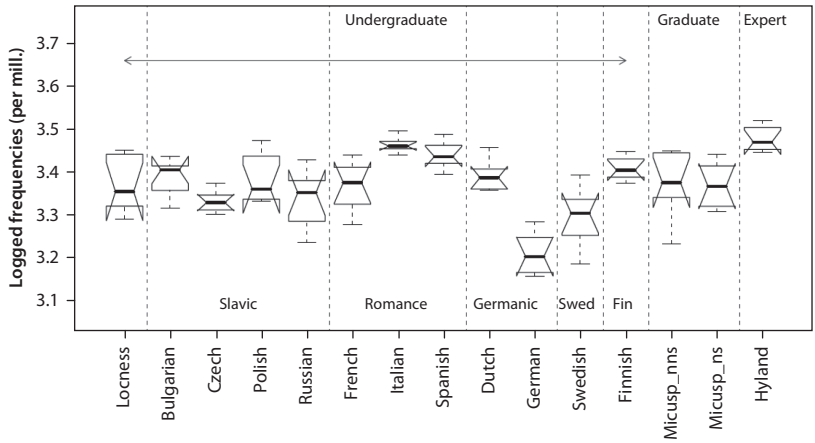

Figure 4. Type frequency of items from AFL (Core and Written) (AFL-grams)

(HYLAND) and all other groups, with no significant differences between these themselves. Thus, for AFL-defined formulas, there are clear effects of high levels of expertise (Expert > A grade graduate ≈ Undergraduate), and no effect of L1/L2 status.

Figure 5 summarizes these patterns for the various operationalizations.

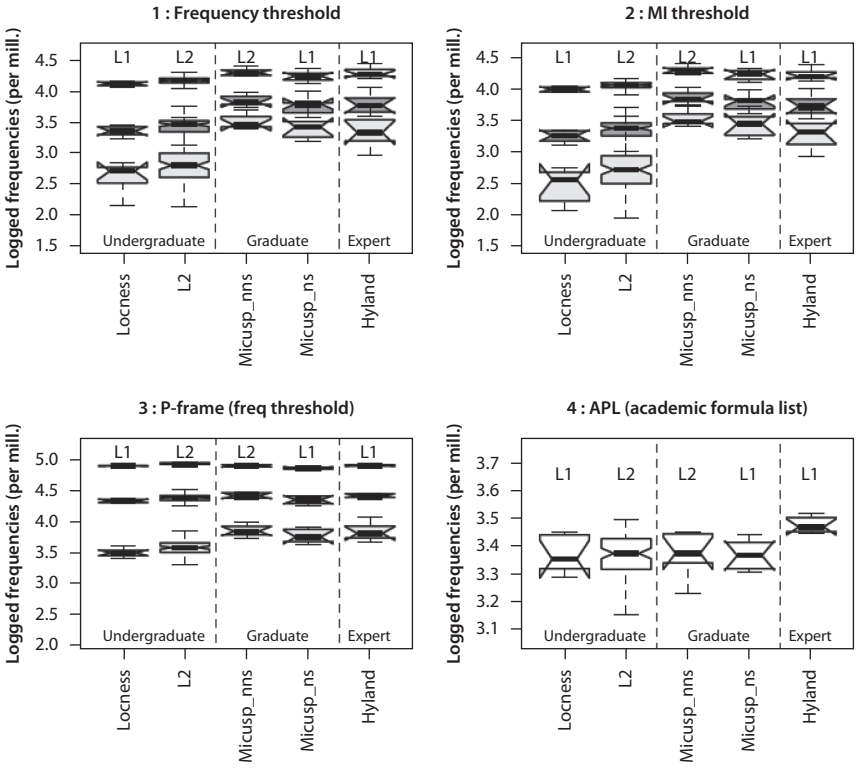

Figure 5. Effects of Expertise and L1/L2 status using the different operationalizations

4. Conclusion

These results demonstrate that different definitions of formulaic language produce different patterns of relationship with expertise and L1/L2 status. We discuss each of these patterns in turn.

Frequency-defined and MI-defined formulas are both more prevalent in advanced writing at expert and graduate levels than in undergraduates. For MI-defined formulas, there are clear effects of expertise (Expert ≈ A-grade Graduate > Undergraduate), but no effect of L1/L2 status. Advanced writing proficiency makes

use of considerable amounts of formulaic language. There are the formulas of the trade – ESP genre-specific formulas such as *nozzle melt pressure*, *drug users who seroconverted*, *under a mistake of law*, *the republic of Somalia*, etc. There are also the conventionalized but wider-ranging EAP rhetorical devices such as *be taken into account*, *due to the fact that*, *it is likely that*, *to some extent*, etc. (Simpson-Vlach & Ellis 2010). These formulas are high in MI. When we define formulaic language in academic writing as being above a threshold MI value, we find a greater density in expert and graduate writing than in undergraduate writing. These formulas are the tools of the trade that differentiate expert from apprentice. It requires considerable experience of disciplinary language to be able to "talk the talk".

For frequency-defined formulas, there are also effects of expertise (Expert ≈ Graduate > Undergraduate), with, if anything, L2 learners producing more formulas than their peers. This later point needs further investigation as there are likely effects of text sampling on the recurrence of formulaic pattern, with the prompt questions driving the more common formulaic sequences in ICLE (e.g. *the opium of the masses*, *the birth of a nation*, *the generation gap*, ICLE French) and LOCNESS (e.g. *the Joy Luck Club*, *in Le Mythe de Sisyphe*, *the root of all evil*). MICUSP (especially MICUSP-NS) reflects common formulaic sequences from reference sections (e.g. *American Journal of Public Health*, *Hispanic Journal of Behavioral Sciences*, *levels of psychological well-being*). HYLAND, with its greater diversity of topics across disciplines, shows less of these sampling foci. It is clear that design choices in the sampling of corpora, such as the particular prompts used, how many participants in the sample attempt the same question, how many texts are selected from the same discipline, whether you include references, etc., have large effects upon the type and amount of formulaic language found.[9] In computing, a phrase of high currency is "garbage-in, garbage out" (GIGO). We could analogize the same for prompts (PIPO), references (RIRO), figure legends (FIFO), dates (DIDO), and the like. Corpus linguistics has developed considerable expertise in corpus composition. But different aims call for different design decisions, and, as we show here, each has an influence on the language composition and, therefore, the formulas caught by different types of net and trawl.

For P-frames, there are neither effects of expertise nor of L1/L2 status. A possible explanation for this lack of significant effects lies in the concept of the p-frame itself. As discussed above, p-frames systematically group n-grams which vary in only one position. They hence provide abstractions of formally (and often semantically) related formulaic sequences. While they help us determine the degree of fixedness of language units, they also require us to look closely at their sets of variants (i.e. the words that most commonly fill the variable position) which enables insights into meaning creation and patterning. This may imply that, if only measured or counted but not manually analysed, p-frames *conceal* the kind

of variation that may tell a novice from an expert writer. So, even if the types and numbers of p-frames in our various datasets are similar, the actual realizations and lists of variants for each p-frame may be very different. For example, the four most frequent p-frame variants of *in the * of* in HYLAND are *case*, *presence*, *context*, and *absence*, whereas the most frequent variants of the same frame in LOCNESS are *case*, *form*, *course*, and *middle*. Incorporating variant types into the measure, for instance by means of a variant/p-frame ratio as suggested by Römer (2010), and a detailed manual analysis of p-frame variants would be required to support this hypothesis. Our observation at this point is that bare numbers of p-frames (separated from their types of variants) are not the most useful measure of formulaicity across text types.

For AFL-defined formulas, there are clear effects of high levels of expertise (Expert > A grade Graduate ≈ Undergraduate), and, again, no effect of L1/L2 status. The expert (HYLAND) authors are senior scholars, who have had multiple-year university training and experience in getting published in peer-reviewed journals. They are clearly differentiated from both the novice academic writers who contributed to ICLE, LOCNESS and those who produced A-grade MICUSP papers on their way to developing expert writing skills and becoming accepted members of academic communities of practice. The fact that there are no effects of L1/L2 status suggest that learning these, for natives and non-natives alike, is akin to learning another language. Clearly it takes a great amount of experience and instruction to become idiomatic in particular specialist genres of English for Academic Purposes.

These demonstrations that different definitions of formulaic language produce different patterns of relationship with expertise and L1/L2 status tell us that methodological choices have weighty consequences. Choices of operationalization entail that different researchers are researching and theorizing different phenomena (see Paquot & Bestgen 2009). Choices of corpus design (the number of participants, the nature of their task and prompts, and the amount of language they produce, etc.) are equally potent determinants of outcome. There remains, therefore, much basic research to be done to assess how these aspects of formulaicity are affected by potential independent variables of concern for control purposes (text length, type token ratio, mean length of utterance, entropy, vocabulary frequency profiles, number of speakers, range of prompts and topics, etc.) and by variables of greater theoretical weight including: potential text variables such as spoken/written genre, potential subject variables such as native vs. second language status, proficiency, education, etc., and potential situational variables such as degree of preparation, rehearsal, and working memory demand. Investigating the measurement of formulaicity in language texts will not benefit corpus linguistics alone, it will prove the necessary foundations for triangulation within cognitive science and our broader understanding of

psycholinguistics, first and second language acquisition, language instruction, and evaluation and testing.

Stubbs & Barth (2003: 84) argued that "corpus data […] make it possible to study how frequent use leads to conventionalized structure". We agree, but our investigations here into conventionalized formulaic language show the problem to be more complex than it might appear at first sight. Much depends on the design of the corpus, the mix of participants, the variability of prompt, task, and genre, and the statistical definition of formulaicity.

Notes

* This paper is dedicated to Michael Stubbs for demonstrating that "[c]orpus study and computational techniques are a cause for confidence that lexical descriptions in the future will provide more accurate and exhaustive documentation about words, and will give access to patterns in the language which are not accessible to unaided human observation. These patterns are probabilistic…" (Stubbs 1995: 51). He caused a lot of trouble, probably all good.

The authors thank Kumud Bihani and Annie Devine for their help in the corpus analyses.

1. The Academic Word List (AWL) was developed by Coxhead in 1998 from a corpus of 3.5 million words of written academic text. It contains 570 word families that account for around 10% of the vocabulary in this corpus (excluding the 2,000 most frequent words in general English). Coxhead claims the specificity of the AWL is demonstrated by comparing this coverage of vocabulary in an academic corpus to other corpora and sub-corpora. For instance, the AWL accounts for only 1.4% of the words in a 3.5 million word corpus of fiction. See Coxhead (2000) for a discussion of the development and evaluation of the AWL.

2. Although the focus of this study and the methods introduced here are primarily quantitative, we would stress the limitations of defining phraseology or formulaic language solely in quantitative terms (see Gries 2008).

3. While not all authors included in the Hyland corpus are native speakers of English, it can be assumed that articles by non-native speakers were checked and corrected by a native speaker before publication. The notion of proficiency and whether it should be measured using internal or external criteria is not without controversy. Recent analysis and classification of the ICLE subcorpora has highlighted the range of proficiency levels within the corpus (see Granger et al. 2009).

4. As previously noted there is a significant range and variance of proficiency level between and among the ICLE subcorpora (see Granger et al. 2009). We do not intend to ignore or dispute this by grouping the eleven L1 components of ICLE1 and considering them together as instances of low level proficiency in academic writing.

5. We did not remove references or footnotes from the texts in the Hyland and MICUSP samples. It might be argued that this could introduce noise into the n-gram data and lead to higher counts for papers, disciplines and corpora that are heavy in such features. In future analysis we would want to explore this possibility. However, other analyses of MICUSP have explored phraseological items, defined as frequency n-grams, and not found the inclusion of references as a problematic confound (Römer 2009b, Ädel & Römer 2012).

6. Given that there are 52 cards in a standard deck it is not possible to have an equal number of cards in each of the eight groups. Instead you would need four groups of six cards and four groups of seven. If, as in blackjack, face cards were given a value of 10 and Ace cards a value of 1, one solution would be:

Group	# of cards	Cards	Value total
1	7	A 2 5 6 9 10 J	43
2	6	3 4 7 8 10 Q	42
3	7	A 2 5 6 9 10 K	43
4	6	3 4 7 8 10 J	42
5	7	A 2 5 6 9 J Q	43
6	6	3 4 7 8 J K	42
7	7	A 2 5 6 9 Q K	43
8	6	3 4 7 8 Q K	42

7. The total token count of 289,234 for ICLE French in Table 3 differs slightly from the one shown in Table 2 (289,699). The totals in Table 2 were generated using the WordList tool in WordSmith Tools Version 5.0 while those shown in Table 2 and used for all corpora to produce the sample groups were generated by a Python script using a simple whitespace tokenizer.

8. The Core AFL consists of 207 items that were found in both spoken and written academic corpora and occurring in at least six out of all nine disciplinary subcorpora at a frequency level of 10 per million. See Simpson-Vlach & Ellis (2010) for a more detailed discussion of the components of the AFL and examples and a functional classification of the contained phrases.

9. Other variables to consider would include the differences between timed (test context) versus untimed essays (term papers) and differences between the type of paper within the corpora of academic writing utilized. MICUSP, for instance, contains a number of paper types (e.g. reports, critiques, argumentative essays, research papers) each with slightly differing communicative functions (see Römer & O'Donnell 2011).

References

Ädel, A. & Römer, U. 2012. "Research on advanced student writing across disciplines and levels: Introducing the Michigan Corpus of Upper-level Student Papers". *International Journal of Corpus Linguistics*, 17 (1), 1–32.

Arnon, I. & Snider, N. 2010. "More than words: Frequency effects for multi-word phrases". *Journal of Memory and Language*, 62 (1), 67–82.

Bannard, C. & Matthews, D. 2008. "Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations". *Psychological Science*, 19 (3), 241–248.

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

Biber, D., Conrad, S. & Cortes, V. 2004. "*If you look at…*: Lexical bundles in university teaching and textbooks". *Applied Linguistics*, 25 (3), 371–405.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Bod, R. 2001. "Sentence memory: Storage vs. computation of frequent sentences". Paper presented at the *CUNY-2001, Philadelphia, Pennsylvania*.

Bybee, J. 2005. "From usage to grammar: The mind's response to repetition". Paper presented at the *Linguistic Society of America, Oakland*.

Cheng, W., Greaves, C. & Warren, M. 2006. "From n-gram to skipgram to concgram". *International Journal of Corpus Linguistics*, 11 (4), 411–433.

Conklin, K. & Schmitt, N. 2008. "Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers?". *Applied Linguistics*, 29 (1), 72–89.

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 (2), 213–238.

Daudaravičius, V. & Marcinkevičienė, R. 2004. "Gravity counts for the boundaries of collocations". *International Journal of Corpus Linguistics*, 9 (2), 321–348.

Durrant, P. & Doherty, A. 2010. "Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming". *Corpus Linguistics and Linguistic Theory*, 6 (2), 125–155.

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. 2002. "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. 2003. "Constructions, chunking, and connectionism: The emergence of second language structure". In C. Doughty & M. H. Long (Eds.), *Handbook of Second Language Acquisition*. Oxford: Blackwell, 33–68.

Ellis, N. C. 2008. "Phraseology: The periphery and the heart of language". In F. Meunier & S. Granger (Eds.), *Phraseology in Language Learning and Teaching*. Amsterdam: John Benjamins, 1–13.

Ellis, N. C. 2012. "Formulaic language and second language acquisition". *Annual Review of Applied Linguistics,* 32, 17–44.

Ellis, N. C. & Cadierno, T. 2009. "Constructing a second language". *Annual Review of Cognitive Linguistics*, 7 (Special section), 111–290.

Ellis, N. C., Frey, E. & Jalkanen, I. 2009. "The psycholinguistic reality of collocation and semantic prosody (1): Lexical access". In U. Römer & R. Schulze (Eds.), *Exploring the Lexis-Grammar Interface.* Amsterdam: John Benjamins, 89–114.

Ellis, N. C. & Simpson-Vlach, R. 2009. "Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education". *Corpus Linguistics and Linguistic Theory*, 5 (1), 61–78.

Ellis, N. C., Simpson-Vlach, R. & Maynard, C. 2008. "Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL". *TESOL Quarterly*, 42 (3), 375–396.

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

Evert, S. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Doctoral dissertation. Stuttgart: University of Stuttgart.

Fletcher, W. H. 2002–2007. *KfNgram*. Annapolis, MD: USNA.
Flowerdew, J. & Peacock, M. (Eds.) 2001. *Research Perspectives on English for Academic Purposes*. Cambridge: Cambridge University Press.
Granger, S. (Ed.) 1998. *Learner English on Computer*. London: Longman.
Granger, S. & Meunier, F. (Eds.) 2008. *Phraseology: An Interdisciplinary Perspective.* Amsterdam: John Benjamins.
Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. 2009. *The International Corpus of Learner English. Version 2. Handbook and CD-Rom*. Presses Universitaires de Louvain: Louvain-la-Neuve.
Gries, St. Th. 2006. "Some proposals towards more rigorous corpus linguistics". *Zeitschrift für Anglistik und Amerikanistik*, 54 (2), 191–202.
Gries, St. Th. 2008. "Phraseology and linguistic theory: A brief survey". In S. Granger & F. Meunier (Eds.), *Phraseology: An Interdisciplinary Perspective.* Amsterdam: John Benjamins, 3–25.
Gries, St. Th. 2013. "Data in construction grammar". In G. Trousdale & T. Hoffmann (Eds.), *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press, 93–108.
Gries, St. Th. & Divjak, D. S. (Eds.) 2012. *Frequency Effects in Cognitive Linguistics (Vol. 1): Statistical Effects in Learnability, Processing and Change*. Berlin: Mouton de Gruyter.
Gries, St. Th. & Stefanowitsch, A. 2004. "Extending collostructional analysis: A corpus-based perspective on 'alternations'". *International Journal of Corpus Linguistics*, 9 (1), 97–129.
Gries, St. Th. & Wulff, S. 2005. "Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora". *Annual Review of Cognitive Linguistics*, 3, 182–200.
Gurevich, O., Johnson, M. A. & Goldberg, A. E. 2010. "Incidental verbatim memory for language". *Language and Cognition*, 2 (1), 45–78.
Guthrie, D., Allison, B., Lin, W., Guthrie, L. & Wilks, Y. 2006. "A closer look at skip-gram modelling". In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC-2006. Genoa, Italy, 2006*, 1222–1225.
Hayes, B. 2002. "The easiest hard problem". *American Scientist*, 90 (3), 113–117.
Hebb, D. O. 1949. *The Organization of Behaviour.* New York: John Wiley & Sons.
Hoey, M. 2004. "The textual priming of lexis". In G. Aston, S. Bernardini & D. Stewart (Eds.), *Corpora and Language Learners*. Amsterdam: John Benjamins, 21–42.
Hoey, M. 2005. *Lexical Priming: A New Theory of Words and Language.* London: Routledge.
Hyland, K. 1998. *Hedging in Scientific Research Articles.* Amsterdam: John Benjamins.
Hyland, K. 2004. *Disciplinary Discourses: Social Interactions in Academic Writing*. 2nd ed. Ann Arbor: University of Michigan Press.
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*. Amsterdam: John Benjamins, 229–254.
Karmarker, N. & Karp, R. M. 1983. *The Differencing Method of Set Partitioning.* Technical Report No. UCB/CSD-83-113. University of California, Berkley.
Leech, G. N. 2011. "Frequency, corpora and language learning". In F. Meunier, S. De Cock, G. Gilquin & M. Paquot (Eds.), *A Taste for Corpora: In Honour of Sylviane Granger*. Amsterdam: John Benjamins, 7–32.
Manning, C. D. & Schütze, H. 1999. *Foundations of Statistical Natural Language Processing.* Cambridge, MA: The MIT Press.
Nattinger, J. R. & DeCarrico, J. 1992. *Lexical Phrases and Language Teaching.* Oxford: Oxford University Press.

O'Donnell, M. B. & Römer, U. In preparation. "Investigating the interaction between phraseological items and textual position".
Oakes, M. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.
Paquot, M. & Bestgen, Y. 2009. "Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction". In A. Jucker, D. Schreier & M. Hundt (Eds.), *Corpora: Pragmatics and Discourse*. Amsterdam: Rodopi, 247–269.
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*. London: Longman, 191–225.
Robinson, P. & Ellis, N. C. (Eds.) 2008. *A Handbook of Cognitive Linguistics and Second Language Acquisition*. London: Routledge.
Römer, U. 2009a. "The inseparability of lexis and grammar: Corpus linguistic perspectives". *Annual Review of Cognitive Linguistics*, 7, 141–163.
Römer, U. 2009b. "English in academia: Does nativeness matter?". *Anglistik: International Journal of English Studies*, 20 (2), 89–100.
Römer, U. 2010. "Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews". *English Text Construction*, 3 (1), 95–119.
Römer, U. & O'Donnell, M. B. 2011. "From student hard drive to web corpus (Part 1). The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP)". *Corpora*, 6 (2), 159–177.
Schmitt, N. (Ed.) 2004. *Formulaic Sequences*. Amsterdam: John Benjamins.
Scott, M. 2008. *WordSmith Tools Version 5.0*. Liverpool: Lexical Analysis Software.
Scott, M. & Tribble, C. 2006. *Textual Patterns. Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.
Simpson-Vlach, R. & Ellis, N. C. 2010. "An Academic Formulas List (AFL)". *Applied Linguistics*, 31 (4), 487–512.
Sinclair, J. McH. 1987. "The nature of the evidence". In: J. McH. Sinclair (Ed.), *Looking Up. An Account of the COBUILD Project in Lexical Computing*. London: HarperCollins, 150–159.
Sinclair, J. McH. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Sinclair, J. McH. 1996. "The search for units of meaning". *Textus*, IX (1), 75–106.
Sinclair, J. McH. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.
Sinclair, J. McH. 2005. "The phrase, the whole phrase, and nothing but the phrase". Paper presented at *Phraseology 2005, Louvain*.
Stubbs, M. 1995. "Collocations and semantic profiles: On the cause of the trouble with quantitative studies". *Functions of Language*, 2 (1), 23–55.
Stubbs, M. 2001. *Words and Phrases. Corpus Studies of Lexical Semantics*. Oxford: Blackwell.
Stubbs, M. 2007. "Quantitative data on multi-word sequences in English: The case of the word *world*". In M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert, *Text, Discourse and Corpora*. London: Continuum, 163–189.
Stubbs, M. & Barth, I. 2003. "Using recurrent phrases as text-type discriminators: A quantitative method and some findings". *Functions of Language*, 10 (1), 61–104.
Swales, J. M. 1990. *Genre Analysis. English in Academic and Research Settings*. Cambridge: Cambridge University Press.
Swales, J. M. 2004. *Research Genres. Exploration and Applications*. 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 (5), 523–534.

Tomasello, M. 2003. *Constructing a Language*. Boston, MA: Harvard University Press.
Tremblay, A., Derwing, B., Libben, G. & Westbury, C. 2012. "Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks". *Language Learning*, 61 (2), 569–613.
Wray, A. 2000. "Formulaic sequences in second language teaching: Principle and practice". *Applied Linguistics*, 21 (4), 463–489.
Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Authors' addresses

Matthew Brook O'Donnell
University of Michigan
Institute for Social Research
426 Thompson St.
Ann Arbor, MI 48106
USA

mbod@umich.edu

Nick C. Ellis
University of Michigan
Department of Psychology
530 Church St.
Ann Arbor, MI 48109
USA

ncellis@umich.edu

Ute Römer
Georgia State University
Department of Applied Linguistics and ESL
34 Peachtree St., Suite 1200
Atlanta, GA 30303
USA

uroemer@gsu.edu