To what extent do native and non-native writers make use of collocations?

PHILIP DURRANT AND NORBERT SCHMITT

Abstract

Usage-based models claim that first language learning is based on the frequency-based analysis of memorised phrases. It is not clear though, whether adult second language learning works in the same way. It has been claimed that non-native language lacks idiomatic formulas, suggesting that learners neglect phrases, focusing instead on orthographic words. While a number of studies challenge the claim that non-native language lacks formulaicity, these studies have two important shortcomings: they fail to take account of appropriate frequency information and they pool the writing of different learners in ways that may mask individual differences. Using methodologies which avoid these problems, this study found that non-native writers rely heavily on high-frequency collocations, but that they underuse less frequent, strongly associated collocations (items which are probably highly salient for native speakers). These findings are consistent with usage-based models of acquisition while accounting for the impression that non-native writing lacks idiomatic phraseology.

1. Introduction

It is becoming increasingly apparent that language is largely formulaic in nature, and that the competent use of formulaic sequences is an important part of fluent and natural language use (Cowie 1998; Nattinger and DeCarrico 1992; Pawley and Syder 1983; Schmitt 2004; Wray 2002). It has also been suggested that formulaic language plays an important role in language acquisition. Following the early lead of child language researchers such as Clark (1974) and Peters (1983), 'usage-based' models of language have recently been developed which see first language learning as a process in which rote-learned, formulaic chunks are gradually subject to analysis and abstraction (Tomasello 2003). Ellis (2003) has proposed that a similar model might be applied to second language acquisition.

IRAL 47 (2009), 157–177
DOI 10.1515/iral.2009.007
0019042X/2009/047-157

One problem with a usage-based approach to second language learning, however, is that it is not yet clear to what extent typical non-native speakers have access to formulaic language. Kjellmer (1990) has made the claim that, unlike natives, who often have the most natural phrase for their meaning pre-constructed and ready-at-hand, even quite advanced learners tend not to know much formulaic language. This forces them to piece structures together word-by-word in ways that they can only hope will prove acceptable – as Kjellmer puts it, their "building material is individual bricks, rather than prefabricated sections" (1990: 124). He claims this is a major reason why otherwise competent non-native speakers can sound unidiomatic. This view appears to constitute a serious challenge to any strong formula-based account of second language learning, suggesting instead a picture similar to that described by Wray, on which adult second language users – influenced perhaps by their more mature, more 'analytical' cognitive abilities – focus from the outset on individual words; only later, and only imperfectly, learning formulaic expressions through conscious effort (Wray 2002: 207–208).

Kjellmer's verdict on non-native language, while intuitively appealing, is presented without empirical support. To adjudicate between 'bottom-up' and 'top-down' views of second language learning we need, therefore, a more thorough investigation of the presence – or absence – of formulaicity in advanced learner language. Some important steps have already been taken in this area. DeCock et al (1998), Oppenheim (2000), Foster (2001), and Adolphs and Durow (2004) have all looked at the use of formulaic language in advanced non-native speech, while Yorio (1989), Granger (1998), Lorenz (1999), Howarth (1998), Kaszubski (2000) and Nesselhauf (2005) have investigated writing. The general picture which has emerged from these studies is that advanced learners do appear to use formulaic language (in some cases quite self-consciously (Oppenheim 2000)), but often not to the same extent as natives (Foster 2001; Granger 1998; Howarth 1998). At the same time, learners tend to overuse (in comparison to native norms) a small range of favourite phrases, especially if they are frequent/neutral items or are cognate to L1 forms (De Cock et al.1998; Foster 2001; Granger 1998; Kaszubski 2000; Lorenz 1999; Nesselhauf 2005).

While these patterns are suggestive, the studies behind them have two important shortcomings. Firstly, most are based on the analysis of corpora consisting of the speech or writing of large numbers of learners (the one exception is Adolphs and Durow's study, whose very small sample of two students presents its own problems). It is not clear, therefore, to what extent their results mask variability between different learners (a point acknowledged by Howarth (1998: 177)). If there are regular and stable norms in the extent to which natives and non-natives make use of formulas, this is not a problem. However, such regularities have not been established for either group. Without estab-

lished norms, and given that variability seems to be the rule in most second language learning and use, the significance of the averaged-out figures which these studies present is not clear.

The second shortcoming of existing studies concerns how formulas are defined and identified. There have been four main approaches here. One has relied on native speaker intuition that a piece of language is formulaic (Foster 2001; Yorio 1989). A second has studied all word combinations of a particular grammatical form (e.g., *-ly* amplifier-adjective combinations, such as *perfectly natural*), regardless of whether they are 'formulaic' in any defined sense (Granger 1998). A third has focused specifically on 'collocations' as they are defined in the so-called 'Russian school' of phraseology (Cowie 1998). In this tradition, collocations are typically identified as those combinations in which either words take on meanings which they do not have in other environments (e.g., *curry favour*) or there are arbitrary restrictions on what words can be substituted into a particular phrase (e.g., the phrase *commit* + [something wrong or illegal] is a collocation because *commit a lie/deceit/delinquency* are arbitrarily blocked) (Howarth 1998; Kaszubski 2000; Nesselhauf 2005). A final approach has based identification on the frequency of occurrence of items within the corpus being studied. Thus, Lorenz looked at intensifier-adverb pairs in parallel native and non-native corpora and used the statistical 'association measures', *t*-score and mutual information, to identify which pairs were strong collocations (Lorenz 1999).

This diversity of approaches is in itself no bad thing, providing us with a fruitful variety of perspectives on the phenomenon. What is problematic, however, is that none of these approaches tells us about what is perhaps the paradigm example of formulaic language – i.e., 'collocations' as they have been defined by corpus linguists of the 'neo-Firthian' school. In this tradition, collocations are characterised as words which appear together in the language more often than their individual frequencies would predict (Hoey 1991; Jones and Sinclair 1974; Kjellmer 1990). Though collocation was originally conceived as a purely textual phenomenon, linguists in the Firthian tradition have come to interpret it in psycholinguistic terms. Frequent collocation is taken to indicate the presence of "semi-preconstructed phrases that constitute single choices" for the language user (Sinclair 1987: 320), or of "a psychological association between words" (Hoey 2005: 5). Presumably, it is these 'semi-preconstructed phrases' or 'psychological associations between words' which second language learners need appropriately to acquire if they are to become native-like users of collocation. Such frequent, psychologically instantiated, collocations will overlap with, but will not be co-extensive with, the 'frozen' or 'semantically anomalous' collocations focused on by 'Russian school' linguists. It remains unclear how accurately and comprehensively frequency-based methods are able to capture such items (though it seems likely

that they would be reasonably successful: difficulties in pinning down the notion of semantic opacity have led Russian school analysts such as Nesselhauf (2005: 29) to rely on the range of free substitutability of elements as their sole criterion in identifying collocation; a frequency-based measure of lexical association such as *mutual information* may retrieve such 'fixed' items quite effectively). However, since many high frequency collocations are neither semantically opaque nor frozen in form, a listing of such items will certainly include items falling outside the Russian school criteria.

The study which comes closest to the Firthian conception is Lorenz (1999), who does use frequency data in identifying collocations. However, because the source of this data is learners' own writing, rather than any broader corpus which might be representative of typical English usage, his analysis does not identify words which are collocations *in English*, but rather words which are commonly associated within learners' own writing. While this approach can give us useful information – in particular, it is able to indicate 'idiosyncratic' combinations which are non-nativelike but nevertheless formulaic for non-natives – it does not (as Lorenz himself takes pains to point out (1999:187)) tell us about learners' use of items which are frequent collocations in English.

Using approaches which do not give frequency information for a language in general has two important drawbacks. First, collocations in the frequency-based sense are prevalent in language and have been hypothesised to be key to the 'naturalness' of native production (Hoey 2005: 2–7; Kjellmer 1990). Approaches which leave this type of collocation out of the picture (and focus exclusively on, for example, 'restricted' or 'semantically anomalous' collocations) run the risk of overlooking a large and important part of formulaic language.

Second, usage-based models hold that a major determining force in the acquisition of formulas is the frequency of occurrence and co-occurrence of linguistic forms in the input (Barlow and Kemmer 2000; Ellis 2003). Since the studies reviewed above (including that of Lorenz) do not provide any information about the frequencies of items in the input, they are unable to comment on this hypothesis in any informed way. The importance of this shortcoming is highlighted by the fact that both Nesselhauf (2005: 224–225) and Lorenz (1999: 181) find themselves proposing frequency-based hypotheses, which they are unable properly to test because they do not have the relevant data. At a more global level, a shortage of frequency-based studies also limits our understanding of the role formulaic language plays in second language acquisition, an area in which our knowledge is currently severely constrained by a lack of empirical data (Ellis 2003: 73).

The present study explores native and non-native use of high-frequency collocations using methodologies which both overcome the problem of 'averaging out' across learners, and take advantage of frequency information from a large

native corpus to focus specifically on collocations which are frequent in English.

2. Methodology

2.1. *Introduction*

This study focuses on English native and non-native writers' use of collocation as it has been defined in the 'frequency-based' tradition represented by such writers as Sinclair (2004), Hoey (2005), and Stubbs (1995). In this approach, collocation refers to "the relationship a lexical item has with items that appear with greater than random probability in its (textual) context" (Hoey 1991: 7). That is, words are collocates if, in a given sample of language, they are found together more often than their individual frequencies would predict (Jones and Sinclair 1974: 19). This definition aims to pick out word pairs whose high frequency indicates a genuine collocational relationship between words, while passing over those which are frequent simply because their constituents are frequent (e.g., *in the; of a*).

2.2. *Native and non-native texts*

This research will compare the collocations found in non-native texts against those found in native texts. The first set of non-native texts used in the study are research assignments produced as project work for courses in English for Academic Purposes (EAP). This text type was chosen because it is one of the few varieties of extended non-native writing. It was thought necessary to use such extended pieces because the study will rely on analysing the extent of collocation use in individual texts and it was suspected that statistically robust trends may only emerge in longer stretches of writing where larger numbers of collocations could be identified. The essays were written by two groups of learners: postgraduate students on pre-sessional EAP courses at a British university; and first-year undergraduates on in-sessional EAP courses at an English-medium university in Turkey.[1] To explore whether the analysis could also work for less extended texts, a set of shorter essays was also analysed. These comprised short compositions written by pre-sessional students at a British university and short 'argumentative' essays from the Bulgarian subcorpus of the International Corpus of Learner English (ICLE) (Granger et al. 2002).

1. Part of this corpus was provided Robin Turner.

Identifying native texts that are equivalent in type to non-native writing is, as other researchers have noted, highly problematic (Granger et al. 2002: 40; Lorenz 1999: 14). The long non-native texts under analysis here do not have readily available native-speaker equivalents: EAP research projects are different in type from normal academic research projects, since they are produced in a class focusing primarily on generic writing and academic skills, without specialist topic-based input, and are intended to be read by an English teacher, rather than by a subject lecturer. In lieu of strictly parallel corpora, therefore, two sets of native writing were analysed which were taken to resemble the EAP projects in different and complementary ways: postgraduate writing (assignments from students on the MA degree in Applied Linguistics at a British university), and essays from the current affairs magazine *Prospect*. The former are similar in form to the EAP projects, but more specialised in topic, since they are written with the support of content-based courses and are intended for an expert readership. The latter are argumentative essays of a similar length to the academic papers. Though distinct in style from academic writing, they are similar to the non-native texts in that they are of similar length, are formal in style, present an argument, and are intended for a general lay audience rather than for specialists.

As a comparison for the shorter non-native texts, two sources were again used. One was argumentative essays written under timed conditions by British undergraduates on the topic, 'A single Europe: A loss of sovereignty for Britain'. These essays were collected by Granger and her colleagues for the Louvain Corpus of Native English Essays (LOCNESS) (Granger et al. 2002: 41) with the specific intention of paralleling texts in ICLE. While these texts are similar in type to the shorter non-native texts, the fact that they are all written on a single topic introduces a risk of skewed data. To incorporate a broader range of topics, opinion articles from two UK newspapers (*The Guardian* and *The Observer*) were also analysed. These short, argumentative pieces are perhaps the closest readily-available parallel to the short compositions produced by the learners.

A total of 96 texts were analysed: 24 long native speaker texts (hereafter referred to as 'NS Long'), 24 long non-native texts ('NNS Long'), 24 short native speaker texts ('NS Short') and 24 short non-native texts ('NNS Short'). Table 1 describes the four sets of texts in detail.

2.3. *Procedure*

2.3.1. *Identification of word combinations.* The present analysis was limited to directly adjacent premodifier-noun word pairs (including both adjective-noun and noun-noun combinations). Modifier-noun combinations were chosen because they were found to be particularly common in the texts analysed, and

Table 1. *Summary of texts analysed in the study*

Type	Sub-type	Description	Number of texts	Number of writers	Total words	Mean words/ text	Writers' L1
NS Long	Prospect	essays from the 'international' section of the current affairs journal *Prospect*	12	12	41304	3442	English
	Academic	academic essays written by students on the MA programme in Applied Linguistics at a British university. 2 essays each were taken from 6 different MA courses	12	7	37429	3119	English
NNS Long	British EAP Project	research projects written by non-native students as part of their final assessment for a pre-sessional course in EAP at a British university. Essay topics are taken from a variety of subject areas, reflecting the academic interests of the students (7 business finance/management; 3 law; 1 classics; 1 political science)	12	12	39145	3262	7 Mandarin 1 Arabic 1 French 1 Greek 1 Korean 1 Russian

Table 1. *Summary of texts analysed in the study*

Type	Sub-type	Description	Number of texts	Number of writers	Total words	Mean words/ text	Writers' L1
	Turkish EAP Project	academic essays written by non-native students for an in-sessional course in EAP during the first year of their degree at an English medium university in Turkey. 6 come from a course based around the themes of the nature-nurture debate and philosophical concepts of personal identity; 6 were from a course based on the philosophy of happiness.	12	12	33217	2768	Turkish
NS Short	Opinion articles	opinion articles from *The Guardian* and *The Observer* newspapers	12	12	8401	700	English
	LOCESS essays	timed essays (1 hour) on the topic of European integration written by British undergraduates	12	12	6734	561	English
NNS Short	British short essays	short compositions written by postgraduate students on a pre-sessional course in English for Academic Purposes at a British university. 6 compositions were on the topic of *Consumerism*; 6 were on the topic of *Education*.	12	6[a]	7936	661	5 Mandarin 1 Russian

Table 1. *Summary of texts analysed in the study*

Type	Sub-type	Description	Number of texts	Number of writers	Total words	Mean words/ text	Writers' L1
	Bulgarian subcorpus of ICLE	short argumentative essays written by students at Sofia University "St Kliment Ohridski". All writers were reported to have spent two years studying English at university level.	12	12	6860	572	Bulgarian

a All of these writers are also represented in 'British EAP Project'.

so provided a rich source of data. Only directly adjacent pairs were used since admitting combinations at a wider range of distances ran the risk of making association measures non-comparable between collocations.

All such pairs were manually extracted from the texts. No attempt was made to filter pairs which might be considered words in their own right (e.g., *prime minister; martial arts*) – such pairings are taken to represent merely one extreme on the scale of collocational fixity.

Combinations were not included if they contained one of the following elements:

– proper nouns (identified by capitalization);
– acronyms defined in the paper (e.g., 'CCT' for 'cross-cultural training')
– pronouns;
– possessives;
– semi-determiners – as listed in Biber et al. (1999), i.e., *same, other, former, latter, last, next, certain, such*;
– numbers/ordinals.

Since the study aims to draw conclusions regarding the performance of the writers themselves, quotations were not included in the analysis.

To keep the calculation of association measures (see below) relatively straightforward, only directly adjacent word pairs were included in the analysis. Thus, where more than one adjective modifies a noun (e.g., *beautiful green eyes*), only the final adjective-noun pair (*green eyes*) is included. Where a premodifying noun is itself premodified, only pairs in the group where the modifier can be read as modifying the succeeding noun itself are included: e.g., from the phrase *national security adviser*, two collocations are extracted, *national security* and *security advisor*; in *local power plant workers, power plant* and *plant workers* are recorded, but not *local power* since *local* doesn't modify *power.*

This procedure retrieved a total of 10,839 word combinations from the 96 texts. The total number of combinations for each text type and the average numbers of combinations retrieved for each text are shown in Table 2. Since different text types were of characteristically different lengths, Table 2 also shows these averages normalised to combinations per 1000 words of text.

2.3.2. *Calculation of collocational strength.* We used two types of frequency-based methods to determine the collocational strength of the extracted word combinations. The first involved simply tallying how frequently each of the combinations occurred in the British National Corpus World Edition (BNC).[2] Since the BNC is one of the largest and most representative corpora

2. The program for extracting frequency data about the target word combinations (i.e., the fre-

Table 2. *Summary of combinations retrieved*

Type	Sub-type	Total combinations retrieved	Average combinations/ text	Average combinations/ 1000 words
NS Long	Prospect	2845	204.25	59.34
	Academic	1500	196.42	62.97
NNS Long	British EAP Project	2451	237.08	72.68
	Turkish EAP Project	2357	125.00	45.16
NS Short	Opinion articles	513	40.42	57.73
	LOCNESS essays	296	24.67	43.96
NNS Short	British EAP short essays	485	42.75	64.64
	Bulgarian subcorpus of ICLE	392	32.67	57.14

of general English currently available, combinations which occur frequently in it are assumed to have common usage in English.

The second method was calculating 'association measures' of collocational strength. Several of these have been proposed as means of identifying word pairs which are collocations in the current frequency-based sense of appearing with greater than random probability (Manning and Schütze 1999 provide an excellent overview). All of these measures work on the principle of comparing the number of times a collocation appears in a corpus with the number of times it would be predicted to appear by chance on the basis of the frequency of its component words. The most widely used of these measures in British lexicography are *t-score* and *mutual information* (MI) (Evert 2004). However, the two association measures tend to emphasise rather different sets of collocations. In particular, whereas rankings based on *t*-scores tend to highlight very frequent collocations (and so are very similar to rankings based on raw frequency), MI tends to give prominence to word pairs which may be less common, but whose component words are not often found apart (Stubbs 1995). Thus, pairs like *good example, long way*, and *hard work* attain high *t*-scores but low MI scores, while pairs like *ultimate arbiter, immortal souls* and *tectonic plates* attain the reverse. With this in mind, we will analyse the word combinations using both association measures, to tap into the kind of information each measure provides.

quency of each word and of each word pair) from the BNC was developed by Jakub Marecek of the University fo Nottingham School of Computer Science and Information Technology. This program did not use lemmatisation or part of speech information in the extraction process.

It has been suggested that a *t*-score of 2 or above and/or a MI score of 3 or above may be taken as indicative of collocation (Hunston 2002; Stubbs 1995). The present study will take these values as minimum conditions for collocation. However, simply dividing combinations into 'collocations' vs. 'non-collocations' on this basis would not be satisfactory, since this would disguise the evident difference between combinations which narrowly pass the threshold (e.g., *remarkable book; sweet child*) and much stronger collocations (*ethnic minorities; global warming*). Combinations will therefore be classified across a scale of collocational strength. This approach of using association measures to grade collocations, rather than simply dividing items into collocates vs. non-collocates accords with the view taken by Manning and Schütze (1999: 166) and by Evert and Krenn (2001), who maintain that association measures are best used to provide ranked lists of collocational strength, rather than to demarcate clear categories. Moreover, by looking at the spread of collocational strength we can get a much more fine-grained view of the data than would be possible on the basis of a simple division of combinations into 'collocations' and 'non-collocations'.

The extracted collocations were divided into 7 bands of *t*-score, as follows:

$$t = 2 - 3.99; t = 4 - 5.99; t = 6 - 7.99; t = 8 - 9.99; t = 10 - 14.99; t = 15 - 19.99; t \geq 20$$

Piloting showed this banding to provide a maximally fine differentiation whilst maintaining a reasonably high number of instances for each level. Similarly, the MI scores were divided into the following bands:

$$\text{MI} = 3 - 3.99; \text{MI} = 4 - 4.99; \text{MI} = 5 - 5.99; \text{MI} = 6 - 6.99; \text{MI} = 7 - 7.99; \text{MI} = 8 - 8.99; \text{MI} = 9 - 9.99; \text{MI} \geq 10$$

Because association measures are thought to be unreliable for low-frequency collocations, and because corpora cannot provide stable evidence for infrequent events (Stubbs 2001), combinations appearing in the BNC fewer than 5 times were not assigned *t*-scores or MI scores (see Results section).

2.3.3. *Group vs. individual scores* As we saw above, previous analyses of native vs. non-native writing have compared native and non-native corpora as wholes. This runs the risk of disguising differences between individual texts, and may therefore potentially produce misleading results. The present analysis aims to overcome this problem by recording results individually for each text and then comparing the four groups of texts using standard inferential statistics, taking each text as an individual case. The difference between this and previous approaches can be understood with an example. The first part of the analysis looks at the proportion of combinations which are rare in English (appearing fewer than 5 times in the BNC). To describe this, a whole-corpus approach

would simply find one set of figures for each of the four sets of texts. On the approach taken here, a figure is instead calculated for each of the 96 texts. An average is then taken for the 24 texts within each type. The advantage of this approach is that we record not only an average figure for each text type, but also the degree of variation between texts. This enables us to use inferential statistics to find whether texts of one type contain a significantly higher percentage of infrequent collocations than those of another. Significant scores on these tests will indicate relative homogeneity within groups and meaningful differences between them.

3. Results

3.1. *Low frequency combinations*

As a first stage in our analysis, we can ask to what extent native and non-native writers make use of combinations which are rare in British English. Figure 1 shows the percentage of combinations used in each set of texts which appear fewer than 5 times (or which fail to appear at all) in the BNC. The mean percentage of combinations falling into this category in the long native texts is 48 %, while for the long non-native texts the figure is 38 %, a substantial and statistically significant difference (NS $M = 48.19$, $SE = 2.14$, NNS $M = 38.87$, $SE = 1.52$, $t(46) = 3.552$, p (two-tailed) $< .001$, $r = .46$). The shorter texts use in general a lower proportion of low-frequency combinations, but show a similar pattern – i.e., low-frequency items are more prevalent in native than in non-native texts, though in this case the difference is not statistically significant (NS $M = 38.14$, $SE = 3.39$, NNS $M = 31.95$, $SE = 2.63$, $t(46) = 1.42$, p (two-tailed) $> .05$, $r = .21$).

3.2. *Strong collocations*

3.2.1. *T-score analysis* The main focus of our study is on the use of 'strong' collocations. An obvious way of analysing the prevalence of such collocations would be to look at the average number used in a given length of text. While this sort of analysis would give an indication of the 'density' of use of strong collocations, it has the disadvantage of confounding the extent to which writers rely on such collocations with the extent to which they use premodifier-noun constructions in general. Thus, as Table 2 indicates, texts from the group 'British EAP Project' use this construction to a much greater extent than do those from the group 'Turkish EAP Project'. Given this, it is not clear whether a finding that strong collocations are more common in the former than in the latter group of texts (as indeed they are) is due to a greater degree of reliance on

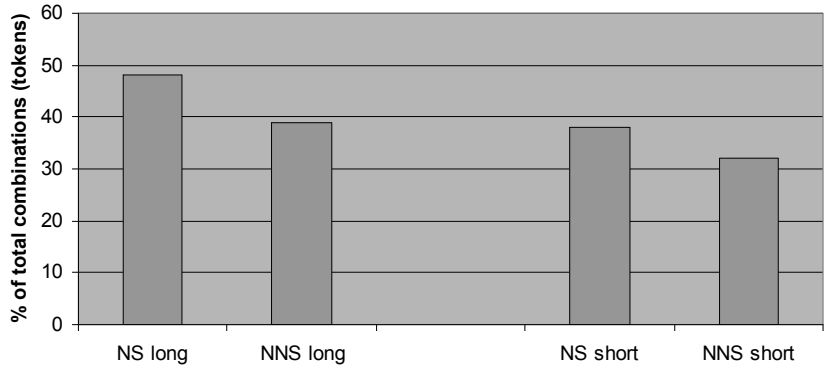

Figure 1. *Mean percentage of combinations which appear < 5 times in BNC*

strong collocations or is merely a product of their greater use of modifier-noun constructions overall. This problem can be overcome by looking not at the total number of collocations used, but rather at the percentage of premodifier-noun combinations which are strong collocations, as indicated by *t*-score and MI statistics. This analysis should give a more valid representation of the degree to which writers rely on conventional collocations.

For each text, the percentage of pre-modifier – noun combinations falling into each *t*-score band was calculated. Figures 2 and 3 summarise the results of this analysis, showing the median percentage of collocation tokens found at each level for each long and short texts respectively (median percentages are used here because the distribution of percentages is not normal within all bands). Since a large number of combinations either appeared in the BNC fewer than 5 times or attained a t-sore of less than 2, the bandings do not sum to 100 %.

Looking first at differences between the longer native and non-native texts, it would appear from Figure 2 that non-native writers take a rather higher proportion of their collocations from the highest bands ($t \geq 10$) than natives. At lower levels, usage appears similar between the two groups of texts. Collapsing the bands into broader 'high' ($t \geq 10$) vs. 'low' ($t < 10$) groupings, enables us to confirm this trend. Non-natives take, on average, 20 % of their collocations from the 'high' band, whereas natives take only 14 %. According to an independent sample t-test, this difference is significant at the $p < .005$ level (NS $M = 13.52$, $SE = 1.42$, NNS $M = 20.16$, $SE = 1.55$, $t(46) = -3.153$, p (two-tailed) $< .005$, $r = -.42$). At the other end of the scale, there is no significant difference between the two sets of texts in their use of the lower

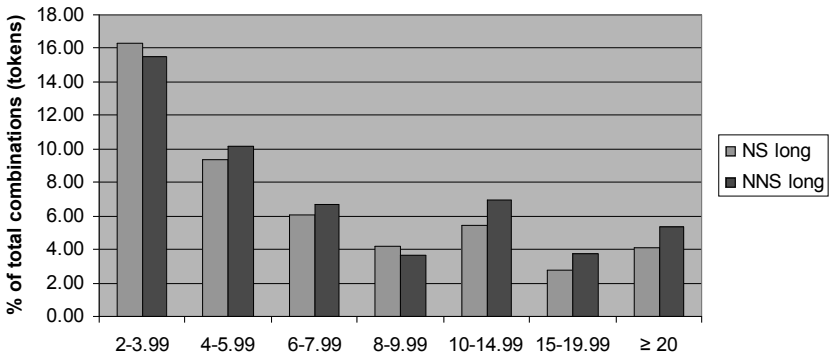

Figure 2. *Median % of collocation (tokens) found at different levels of* t-*score for long texts*

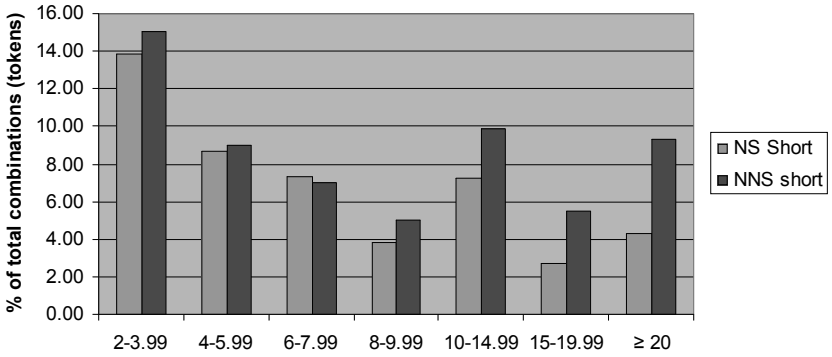

Figure 3. *Median % of collocation (tokens) found at different levels of* t-*score for short texts*

strength collocations (NS $M = 35.69$, $SE = 1.27$, NNS $M = 37.14$, $t(46) = -0.796$, p (two-tailed) $> .05$, $r = .12$).

We have seen that some researchers have claimed that non-native writing is characterised by the repeated use of a small repertoire of collocations (Granger 1998; Kaszubski 2000; Lorenz 1999). That the non-native texts in our data make greater use of repetition than the natives can be confirmed by calculating a collocational type-token ratio (calculated as the mean number of collocation types per 100 collocation tokens) for each text. The median ratio for long native texts is 90, compared with 63 for non-natives. The median ratio for short native texts is 96, compared with 90 for non-natives (note that type-token ratios are typically higher for shorter texts (Richards 1987)). It may be then, that the non-native writers' comparative 'overuse' of strong collocations comes about

because they rely on repeating a few favoured formulas. To check whether this is the case, we can recalculate our data using collocation types rather than collocation tokens. Such an analysis can be interpreted as telling us about the repertoire of collocations demonstrated by each writer.

Using these data to re-examine the differences described above, we find that the pattern of non-native overuse is indeed weakened somewhat. In this case, non-natives continued to take a higher proportion of their collocations from the $t \geq 10$ band than natives but the difference is now much smaller and marginally nonsignificant (NS $Mdn = 11.70$, NNS $Mdn = 14.26$, $U = 200.00$, p (two-tailed) $= .07$, $r = -.26$; non-parametric tests are used because results were not normally distributed within the long non-native texts). Any non-native overuse of the strongest collocations may therefore be the result of the repeated use of favoured items. However, even when repetition is removed from the data, it is fairly clear that non-natives make no less use of strong collocations than natives.

Turning now to the shorter texts, Figure 3 seems to indicate a pattern similar to that seen for natives vs. non-natives as a whole – i.e. relative overuse by non-natives at the higher levels. Again collapsing the results into high ($t \geq 10$) vs. low ($t < 10$) bands, we find significant overuse of high scoring combinations by non-native speakers (NS $Mdn = 18.34$, NNS $Mdn = 26.60$, $U = 190.00$, p (two-tailed) < 0.05, $r = -.29$; non-parametric tests are used because results for short native speaker texts were not normally distributed). Again, the difference is weakened if we look at collocation types rather than tokens (NS $M = 18.95$, $SE = 2.44$, NNS $M = 22.81$, $SE = 1.51$, $t(46) = -1.345$, p (two-tailed) > 0.05, $r = .19$).

3.2.2. *Mutual information analysis.* Mutual information is known to emphasise a rather different set of collocations from *t*-scores, so we also carried out a similar analysis using the MI procedure. Figures 4 and 5 summarise the results of this analysis, showing the median percentage of collocation tokens found at each level for long and short texts respectively.

Again we can start by looking at the differences between the longer native and non-native texts. Reversing the results seen for the *t*-score analysis, Figures 3 and 4 appear to indicate that non-native writers relied to a lesser extent on very strong collocations than did natives. In particular, non-natives show a consistent pattern of 'underuse' at all levels in which MI ≥ 7. An independent samples t-test shows the difference between native and non-native use of MI ≥ 7 collocation tokens not to be significant (NS $M = 17.48$, $SE = 1.30$, NNS $M = 14.95$, $SE = 1.46$, $t(46) = 1.289$, p (two-tailed) $> .05$, $r = .19$). However, if we look at the percentage of collocation *types* taken from these levels, the difference becomes highly significant (NS $M = 15.47$, $SE = 1.00$,

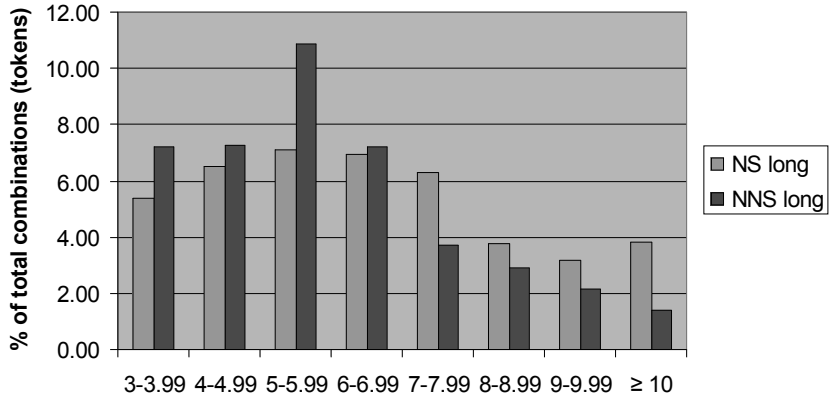

Figure 4. *Median % of collocations (tokens) found at different levels of MI for long texts*

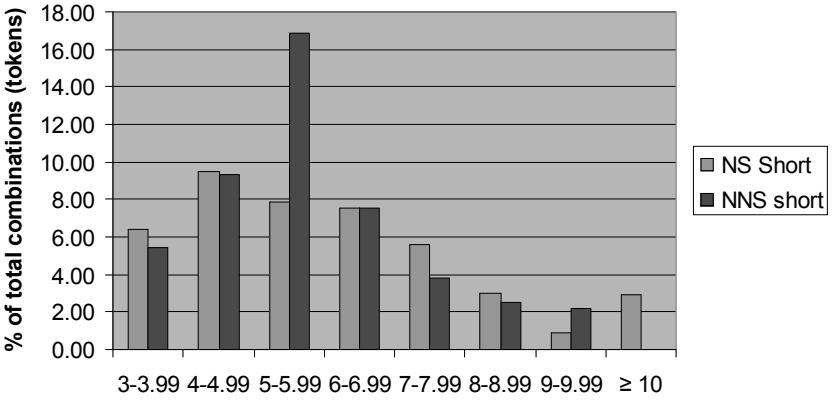

Figure 5. *Median % of collocations (tokens) found at different levels of MI for short texts*

NNS $M = 11.07$, $t(46) = 3.386$, p (two-tailed) $< .001$, $r = .45$). As before then, non-native use of the stronger collocations seems to have been boosted by repetition. Taking a slightly more exclusive band of strong collocations (MI ≥ 8), the difference between the two sets of texts is more emphatic: non-natives show significant underuse of items from these bands in both the analysis by tokens (NS $Mdn = 11.08$, NNS $Mdn = 8.32$, $U = 184.50$, p (two-tailed) $< .05$, $r = -.31$; non-parametric tests are used because results for long non-native speaker texts were not normally distributed) and that by types (NS $M = 9.75$, $SE = 0.71$, NNS $Mdn = 5.85$, $SE = 0.58$, $t(46) = 4.236$, p (two-tailed) $< .001$, $r = .53$).

The shorter texts exhibit a similar, though slightly less robust, pattern. Non-natives show a nonsignificant underuse of strong collocation tokens (MI $\geq$ 8) in comparison to native norms (NS $Mdn = 11.62$, NNS $Mdn = 6.29$, $U = 218.5$, p (two-tailed) $> .05$, $r = -.21$; non-parametric tests are used because results for short non-native speaker texts were not normally distributed), but this difference reaches significance in the analysis of types (NS $M = 11.43$, $SE = 1.30$, NNS $M = 7.88$, $SE = 1.01$, $t(46) = 2.159$, p (two-tailed) $< .05$, $r = .30$).

4. Discussion

This study has aimed to describe the extent to which non-native writers make use of word combinations, and particularly strong collocations, in comparison to native speaker norms, by using methodologies which take advantage of frequency information, and which take account of individual variability between texts. Three main findings have emerged. Firstly, native writers use more low-frequency combinations than non-natives. This trend appears to be fairly consistent across texts, even though it was statistically significant only in the comparison of longer texts. Secondly, non-native writers make at least as much use of collocations with very high *t*-scores as do natives. Since non-natives also tend to repeat certain favoured collocations, if we consider collocation tokens, rather than types, they show a significant overuse of these strong collocations in comparison to native norms. Thirdly, non-native writers significantly underuse collocations with high mutual information scores in comparison with native norms. Again, the repetition of favoured items bolsters the non-native count somewhat, so the difference is more marked on an analysis of collocation types. All of these regularities were less marked in shorter texts, but even here we found sufficient consistency of usage for the same tendencies to emerge, if not always with statistical significance.

How then should we interpret this pattern of results? Firstly, non-natives' relative 'underuse' of low frequency and novel combinations would appear to indicate a degree of conservatism in their production – learners seem to overrely on forms which are (according to BNC data) common in the language. In particular, their extensive use of collocations with very high *t*-scores indicates a preference for very frequent collocations. This conservatism is also indicated by learners' tendency to repeat favoured items. These findings would appear to undermine the idea that non-native writers work in a primarily 'bottom-up' direction from words to phrases. It would seem that second language learners do acquire quite effectively much of the high-frequency phraseology of the target language.

At the same time, however, Kjellmer's intuition that there is something missing from the phraseology of non-native writing does appear to have some truth

to it. In particular, collocations with high mutual information scores (i.e., those which are relatively 'exclusive' to one another, including less frequent collocations) seem to be underused. This is an intuitively satisfying result: learners are quick to pick up highly frequent collocations, but less common, strongly associated items (e.g., *densely populated, bated breath, preconceived notions*) take longer to acquire. This trend for overuse of high frequency items and underuse of high MI items tallies with some previous research. On the basis of his study of native vs. non-native writing, Lorenz (1999: 181) speculates that learners may rely on "attestedly viable, recurrent collocations" , while natives prefer the less frequent, but more strongly-associated pairs characterised by high MI-scores. As Lorenz acknowledges, his method of analysing the data (which relied on the use of frequency counts *within* the non-native corpus, rather than on the use of a reference corpus) was not able to provide a confirmation of this suspicion. The present paper provides the confirmation he needed. Coming from a rather different angle, Ellis and his colleagues (Ellis et al. 2008) provide psycholinguistic evidence for the importance of high frequency collocations to non-native speakers and the importance of collocations with high MI scores to natives: while native speaker speed of processing of 'academic phrases' – as measured by the time taken to recognise phrases and by the time taken to pronounce them – is most strongly predicted by phrases' MI scores, speed of non-native processing of the same items is best predicted by their frequency.

Kjellmer's characterisation of non-native language is, therefore, not entirely wrong, but it is in need of reformulation. Advanced non-native phraseology differs from that of natives not because it avoids formulaic language altogether but because it overuses high-frequency collocations and underuses the lower-frequency, but strongly-associated, pairs characterised by high mutual information scores. Since the latter sort appear (intuitively, and on the psycholinguistic evidence presented by Ellis et al) to be highly salient for native speakers, their absence may be what creates the feeling that non-native writing lacks 'idiomaticity'. However, it is not necessary to posit any radically different ('bottom-up') L2 learning mechanism to explain this absence – their characteristically low frequency of occurrence simply means that such collocations are likely to be acquired later than other parts of nativelike phraseology. This would seem to be quite consistent with a usage-based model of second language learning. Such a pattern does suggest, however, that language teachers wishing to hasten their students along the route of developing an 'authentic' nativelike phraseology may benefit from drawing their attention to collocations attaining high mutual information scores in a corpus of the target language.

Bilkent University
<durrant.phil@googlemail.com>
University of Nottingham
<norbert.schmitt@nottingham.ac.uk>

References

Adolphs, Svenja and Valerie Durow (2004). Social-cultural integration and the development of formulaic sequences. In *Formulaic Sequences: Acquisition, Processing and Use*, Norbert Schmitt (ed.), 107–126. Amsterdam: John Benjamins Publishing Company.

Barlow, Michael and Suzanne Kemmer (eds.) (2000). *Usage-based Models of Language*. Stanford, CA: CSLI Publications.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Clark, Ruth (1974). Performing without competence. *Journal of Child Language* 1 (1): 1–10.

Cowie, Anthony P. (1998). Introduction. In *Phraseology: Theory, Analysis, and Applications*, Anthony P. Cowie (ed.), 1–20. Oxford: Oxford University Press.

Cowie, Anthony P. (ed.) (1998). *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press.

De Cock, Sylvie, Sylviane Granger, Geoffrey Leech and Tony McEnery (1998). An automated approach to the phrasicon on EFL learners. In *Learner English on Computer*, Sylviane Granger (ed.), 67–79. London: Addison Wesley Longman.

Ellis, Nick C. (2003). Constructions, chunking, and connectionism: The emergence of second language structure. In *The Handbook of Second Language Acquisition*, Catherine J. Doughty and Michael H. Long (eds.), 63–103. Oxford: Blackwell.

Ellis, Nick C., Rita Simpson-Vlach and Carson Maynard (2008). Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics and TESOL. *TESOL Quarterly* 41 (3): 375–396.

Evert, Stefan (2004). Computational approaches to collocations. www.collocations.de. Retrieved on: 14 December.

Evert, Stefan and Brigitte Krenn (2001). Methods for the qualitative evaluations of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 188–195. Toulouse, France.

Foster, Pauline (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing*, Martin Bygate, Peter Skehan and Merrill Swain (eds.), 75–94. London: Longman.

Granger, Sylviane (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In *Phraseology: Theory, Analysis, and Applications*, Anthony P. Cowie (ed.), 145–160. Oxford: Oxford University Press.

Granger, Sylviane, Estelle Dagneaux and Fanny Meunier (2002). *International Corpus of Learner English*. Louvain: UCL Presses Universitaires de Louvain.

Hoey, Michael (1991). *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Hoey, Michael (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Howarth, Peter (1998). Phraseology and second language proficiency. *Applied Linguistics* 19 (1): 24–44.

Howarth, Peter (1998). The phraseology of learners' academic writing. In *Phraseology: Theory, Analysis, and Applications*, Anthony P. Cowie (ed.), 161–186. Oxford: Oxford University Press.

Hunston, Susan (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Jones, Susan and John McH. Sinclair (1974). English lexical collocations. A study in computational linguistics. *Cahiers de lexicologie* 24: 15–61.

Kaszubski, Przemek (2000). *Selected Aspects of Lexicon, Phraseology and Style in the Writing of Polish Advanced Learners of English: A Contrastive, Corpus-based Approach*. Poznán: Adam Mickiewicz University.

Kjellmer, Göran (1990). A mint of phrases. In *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, Karin Aijmer and Bengt Altenberg (eds.), 111–127. London: Longman.

Lorenz, Gunter (1999). *Adjective Intensification – Learners Versus Native Speakers: A Corpus Study of Argumentative Writing*. Amsterdam: Rodopi.

Manning, Christopher D. and Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Nattinger, James R. and Jeanette S. DeCarrico (1992). *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

Nesselhauf, Nadja (2005). *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.

Oppenheim, Nancy (2000). The importance of recurrent sequences for nonnative speaker fluency and cognition. In *Perspectives on Fluency*, Heidi Riggenbach (ed.), 220–240. Ann Arbor: University of Michigan Press.

Pawley, Andrew and Frances Hodgetts Syder (1983). Two puzzles for linguistic theory: Native-like selection and nativelike fluency. In *Language and Communication*, Jack C. Richards and Richard W. Schmidt (eds.), 191–226. New York: Longman.

Peters, Ann M. (1983). *The Units of Language Acquisition*. Cambridge: Cambridge University Press.

Richards, Brian (1987). Type/token ratios: What do they really tell us? *Journal of Child Language* 14: 201–209.

Schmitt, Norbert (ed.) (2004). *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins.

Sinclair, John McH. (1987). Collocation: A progress report. In *Language Topics: Essays in Honour of Michael Halliday*, Ross Steele and Terry Threadgold (eds.), 319–331. Amsterdam: John Benjamins.

Sinclair, John McH. (2004). *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Stubbs, Michael (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative methods. *Functions of language* 2 (1): 1–33.

Stubbs, Michael (2001). Texts, corpora, and problems of interpretation: A response to Widdowson. *Applied Linguistics* 22 (2): 149–172.

Tomasello, Michael (2003). *Constructing a Language: A Usage-based Theory of Language Acquisition*. Cambridge MA: Harvard University Press.

Wray, Alison (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Yorio, Carlos A. (1989). Idiomaticity as an indicator of second language proficiency. In *Bilingualism across the Lifespan*, Kenneth Hyltenstam and Loraine K. Obler (eds.), 55–72. Cambridge: Cambridge University Press.