International Journal of Applied Linguistics ◆ Vol. 27 ◆ No. 1 ◆ 2017

Transitional probability predicts native and non-native use of formulaic sequences

Randy Appel and Pavel Trofimovich *Concordia University, Montreal*

Formulaic sequences (FSs), or prefabricated multi-word structures (e.g. on the other hand), are often difficult to identify objectively, and current corpus-driven methods yield structurally incomplete, overlapping, or overly extended structures of questionable psychological validity and pedagogical usefulness. To address these limitations, this study evaluated transitional probability as a potential metric to improve the identification of FSs by presenting 100 four-word sequences from the British National Corpus, varying in transitional probabilities between words, to native and non-native speakers of English (*N* = 293) in a sequence completion task (e.g. for the sake__). Results revealed that the application of transitional probability reduces many of the problems associated with current approaches to FS identification and can produce lists of FSs that are more functionally salient and psychologically valid.

Keywords: formulaic sequences, formulaic language, lexical bundles, n-grams, corpus-driven research

Les expressions stéréotypées (ES), ou les séquences préfabriquées (par exemple, *on the other hand*) sont souvent difficiles à identifier objectivement et les méthodes actuelles basées sur des corpus produisent des structures incomplètes, se chevauchant, ou excessivement étendues, ce qui remet en question leur validité psychologique et leur utilité pédagogique. Pour pallier ces limites, cette étude a évalué le potentiel d'une métrique basée sur la probabilité de transition dans le but d'améliorer l'identification des ES. Pour cela, 100 séquences de quatre mots tirées du *British National Corpus*, variant en probabilité de transition entre les mots, ont été présentées à des locuteurs natifs et non natifs en anglais (n = 293) lors d'une tâche consistant à compléter les séquences (par exemple, *for the sake*__). Les résultats ont révélé que l'application de la probabilité de transition circonscrit plusieurs des problèmes associés aux approches actuelles d'identification de ES, et peut produire des listes de ES plus fonctionnellement saillantes et psychologiquement valides.

Mots clés: expressions stéréotypées, linguistique de corpus, psycholinguistique

doi: 10.1111/ijal.12100

Introduction

The study of formulaic language has been approached from a variety of perspectives following Firth's (1935) assertion that words tend to co-occur in particular patterns. It is now a well-known fact that spoken and written language includes predictable multiword units carrying specific meanings and that these units form a core component of natural language use (Schmitt 2010). According to usage-based approaches to language learning (e.g. Barlow and Kemmer 2000), recurrent units of meaning are represented by usage events or constructions, which refer to pairings of form and meaning (e.g. *I don't know*, *kick the bucket*), and it is the exposure, categorization, and subsequent probability assessments of these events over time that lead to language learning. As exposure to usage events increases, through input and output, probabilities related to the acceptability of utterances are accumulated and eventually used to interpret and produce discourse. Over time, frequent sequences can come to be stored as prefabricated units that are pulled from memory fully formed at the time of use. These structures, referred to as formulaic sequences (FSs), make up a large portion of discourse (Erman and Warren 2000; Schmitt and Carter 2004), contribute to fluent, nativelike speech (Pawley and Syder 1983; Kuiper 1996), characterize proficient language use (Bamberg 1983; McCully 1985), and confer processing advantages in comprehension and production (Peters 1983; Tremblay, Derwing, Libben and Westbury 2011).

With an increased recognition of FSs, research targeting their identification has also grown. However, since formulaic language comes in many forms with varying lengths, degrees of fixedness, levels of grammatical acceptability, and semantic opaqueness, there is no single definition of formulaicity. As a result, researchers have developed different criteria, such as frequency statistics and association indexes, and created a variety of terms (e.g. chunks, amalgams, prefabricated routines) to label the object of study. Thus, terminology is difficult to reconcile across studies, largely because different terms and criteria are applied to the same general concept. Among the most easily recognizable FSs are idioms, such as *a stitch in time saves nine* and *kick the bucket*, due to their semantic opaqueness and non-compositionality. However, other frequent fixed-form FSs, such as *on the other hand*, *by the way*, and *the fact that*, are difficult to label consistently because they serve various discourse functions, are often more semantically clear and compositional, and can be identified according to various criteria.

The goal of the current study was to evaluate a novel corpus-derived criterion for improving the identification of fixed-form FSs, such as *on the other hand* and *by the way*, with the aim of more accurately identifying psychologically valid sequences that are more functionally salient and teacher friendly. As discussed below, current corpus-driven methods yield structurally incomplete, overlapping, or overly extended structures of questionable psycholinguistic validity, functional salience, and pedagogical usefulness (Nekrasova 2009; Simpson-Vlach and Ellis 2010; Liu 2012). Therefore, it is important to develop and test new methods of objectively identifying these structures. Doing so will enable us to understand recurrent word patterns present in various corpora that may otherwise have gone unnoticed. Simply relying on personal reflection or subjective judgements of formulaicity is insufficient in this regard, since many recurrent word patterns only begin to emerge thorough analyses of large amounts of text. Developing accurate methods of identifying FSs is also important because FSs can facilitate language acquisition. Indeed, FSs represent a large portion of native-speaker discourse (Erman and Warren 2000) and focused instruction targeting FSs may help learners incorporate them into their linguistic repertoires (Boers, Eyckmans, Kappel, Stengers and Demecheleer 2006). Thus, to identify fixed-form FSs and evaluate their use, this study targeted the statistical measure of transitional probability, a previously unused metric in this field. The assumption was that transitional probability, which assesses word association strength to indicate utterance boundaries, should lead to more accurate FS identification (i.e. with fewer incomplete, overlapping, and overly extended structures) and should better predict FS use in both native and non-native users, compared to other criteria, such as frequency and mutual information statistics.

Corpus-driven research into identification of FSs

Corpus-driven research, with its focus on large repositories of spoken or written language, has helped to remove much of the subjectivity associated with native-speaker judgements, used previously to define FSs (Conrad 2000). This research largely targets two separate yet associated approaches, namely, *n-grams*, identified according to frequency of occurrence, and *lexical bundles*, identified using frequency and range. However, both methods assume that frequency dictates importance of the structure, and this assumption is illustrated through research on lexical bundles, an increasingly popular method of identifying FSs in various genres and registers.

Defined as “the most frequently recurring sequences of words” (Biber and Barbieri 2007: 264), lexical bundles refer to a subset of formulaic language derived through frequency and range requirements. Identification of lexical bundles involves the analysis of digitized text, sourced from written or oral discourse, through concordancing software (e.g. WordSmith, Collocate) which scans the text for repeated structures. With the size of the ‘window’ (sequence length) usually set at four words (Cortes 2006; Hyland 2008; Biber and Barbieri 2007), the software scans the text beginning with the first word in the corpus. As the window moves forward, it takes four-word ‘pictures’ of the corpus in a progressive manner, such that words 1–4 represent the first sequence followed by words 2–5, 3–6, 4–7 and so on. Each sequence is recorded, and repetitions are tallied to create a list of recurrent sequences.

With minimum frequency threshold often set at 20–40 occurrences per million words, the software can produce lists of sequences meeting this criterion. Generated sequences are subsequently reviewed to ensure adherence to a minimum range requirement, typically set at 3–5 texts (e.g. Biber and Barbieri 2007) or 10% of texts used in the corpus (e.g. Hyland 2008), to eliminate frequent sequences confined to limited texts produced by individual speakers or writers.

The lexical bundle methodology also carries limitations that result in overlapping, overly extended, and structurally or semantically incomplete sequences lacking psychological salience (Nekrasova 2009; Simpson-Vlach and Ellis 2010; Liu 2012). Because this approach relies almost exclusively on the frequency criterion, it often identifies repeated structures that carry little actual meaning. For example, two frequent FSs that appear together in discourse (e.g. *due to* and *the fact that*) will often be presented by the concordancing software as multiple four-word entries with no unitary semantic status (i.e. *due to the fact*, *to the fact that*, *the fact that the*). Because sequence length is determined *a priori* by the researcher, lexical profiling also generates lists which misrepresent complete structural units, such as sequences that cross syntactic boundaries, often with a determiner (*a*/*the*) in terminal position (e.g. *the fact that the*, *the nature of the*). In these cases, the determiner likely belongs to a separate unit of meaning and should not be included with the structure. The often incomplete semantic and structural status of lexical bundles also contributes to difficulties with the assignment of functional roles to them. Cortes (2004), for instance, classifies *the fact that the* and *the nature of the* as discourse organizing bundles. However, Biber, Conrad, and Cortes (2004) view them as serving stance and referential purposes, likely as a result of the utterance-final *the* referring to a new unit of meaning.

Additional problems with lexical bundles relate to sequence length. With lexical searches often limited to four-word units, the assumption is that all FSs include four words. For instance, it is often argued that four-word bundles "offer a clearer range of structures and functions than 3-word bundles" (Hyland 2008: 8) and that four-word sequences subsume three-word units (Cortes 2004). However, as shown above, the functions of many four-word bundles do not seem as clear, and the fact that four-word structures contain three-word units is a drawback, not an advantage. Finally, problems with the identification of lexical bundles also question their psychological status as prefabricated units. Since lexical bundles are identified based on frequency, they often cross syntactic and semantic boundaries and lack clear meanings (i.e. *and one of the*, *going to be the*), which would not normally be associated with usage events in the mind of a language user. This also raises questions about the utility of lexical bundles as pedagogical tools since it would be hard for learners to use FSs which lack clear functional roles.

To sidestep some of the limitations of the lexical bundle approach, researchers recently proposed another statistical measure – termed mutual information index – as a supplement to existing frequency and range criteria

(Simpson-Vlach and Ellis 2010). Mutual information refers to the ratio of the observed frequency of an entire *n*-word sequence in a corpus (e.g. *cup of tea*) relative to the expected frequency of that same sequence occurring by chance alone. Mutual information encompasses probability values for each constituent word in the target structure, with the total probability being a product of all individual word probabilities. The assumption underlying the mutual information statistic is that frequency alone often fails to identify important word associations and that mutual information, with its focus on associations between words, can yield more functionally salient and structurally complete FSs. For example, Simpson-Vlach and Ellis (2010) successfully used mutual information to improve functional salience and significantly reduce the number of identified structures with a determiner in terminal position.

Unfortunately, mutual information does not take into consideration word order, so it can only be used to calculate co-occurrence of an entire sequence, not sequential probability. For instance, the mutual information values for *cup of tea*, *tea of cup*, *cup tea of*, and *tea cup of* are all identical, illustrating the main limitation of this measure, namely, its insensitivity to sequential probabilities of occurrence between elements. Additionally, as identified by Biber (2009), mutual information tends to disfavour sequences that contain high-frequency function words, such as *the*, *of*, *a*, with the consequence that mutual information values become disproportionately reduced for target sequences with highly-frequent words. Therefore, although mutual information has been used to improve some aspects of corpus-driven identification of FSs, more effective approaches are still needed.

Transitional probability

With the overall goal of developing a more effective approach to corpus-driven identification of FSs, this study targeted the measure of transitional probability which has been used in psycholinguistic research on word segmentation and statistical learning (e.g. Aslin and Newport 2012; Mirman, Estes, and Magnuson 2010). Transitional probability is a measure of co-occurrence of segments, syllables, or words in a sequence, estimating the likelihood of a particular element being followed by another. Forward transitional probability, the likelihood of X being followed by Y, establishes the frequency of XY relative to all occurrences of the initial element in the sequence. Similarly, backward transitional probability, the likelihood of X preceding Y, denotes the frequency of XY relative to all instances of the final element in this sequence. In word segmentation research, high transitional probability between syllables suggests that syllables co-occur and likely represent a word-like unit, while low transitional probability between syllables implies word boundaries. Both children and adults can compute such statistics after about two minutes of exposure to a novel sequence, demonstrating above-chance segmentation performance (Aslin and Newport 2012). This study extends this notion from research on speech segmentation to research on formulaic language, to more accurately isolate utterance boundaries and ultimately achieve more precise FS identification.

While similar to mutual information in that it can be used to compare probabilities of occurrence relative to the frequencies of individual elements, transitional probability holds the advantage of taking into account sequence order when making these calculations and can be used to measure the strength of association at different points in a structure. This is an important benefit, suggesting that transitional probability can be combined with standard lexical bundle and *n-gram* methodologies to better understand the boundaries of FSs and thus more accurately identify which sequences function as fixed-meaning units, helping reduce the incidence of overlapping, incomplete, and overly extended structures identified as FSs.

The current study

In light of the numerous limitations associated with the traditional lexical bundle approach and the more recent introduction and use of mutual information statistics, this study evaluated transitional probability as a potentially successful new method of improving the identification of fixed-form FSs in large corpora. This study, whose goal was to test the effectiveness of transitional probability as a metric of FS status with both native English users and second language (L2) learners (i.e. individuals for whom psychologically valid and pedagogically useful FSs will arguably be most relevant), was guided by two closely related research questions:

1. Can the application of forward and backward transitional probability be used to improve the identification of functionally salient and psychologically valid formulaic sequences that are better suited for pedagogical purposes?
2. To what extent does transitional probability, compared to the traditional frequency of occurrence and mutual information statistics, predict native English users' and L2 learners' completion of highly frequent four-word sequences?

The answer to these questions will reveal a greater understanding of how FSs function within discourse and help improve our ability to objectively identify these structures in large corpora. Although a limited range of FSs have begun to appear in some materials aimed at L2 English users, EAP materials often underrepresent this aspect of language and do not provide L2 users with the kinds of FSs they are likely to face in their academic careers (Wood and Appel 2014). It is therefore important to investigate methods of improving the identification of pedagogically useful FSs that can be incorporated into language teaching materials aimed at a variety of L2 English users' needs. To evaluate the effectiveness of transitional probability in the identification of FSs, 100 frequently occurring four-word sequences were extracted from the British National Corpus, with the idea that some represented 'fixed' four-word FSs but others encompassed partial or incomplete structures that are misidentified by applying the traditional lexical bundle approach. These sequences were presented to 138 native English speakers and 155 L2 learners in a sequence completion task, with the fourth word deleted, to examine the rate and variability in sequence completion. Using both native and non-native English users was crucial in testing the utility of transitional probability for research and teaching purposes, on the assumption that the FSs identified through transitional probability should ultimately be useful for L2 learners, who (as argued above) might benefit from psychologically and pedagogically valid FSs in language learning. These scores were tested against the transitional probability, mutual information, and traditional frequency of occurrence statistics to determine which measure best predicted language users' performance.

Method

Participants

The participants included 138 native English speakers and 155 L2 English learners. The native speakers (81 female, 57 male), who were on average 37.9 years old (20–75), resided in the United States because recruiting native English speakers with no multilingual experience and little knowledge of additional languages was problematic in Montreal, Canada, a bilingual French-English city with a large multilingual population. Therefore, native speakers were recruited through online media and tested in a timed, on-line setting. These participants all identified English as their native language, with several reporting basic knowledge of Spanish (7), French (3), Vietnamese (2), American Sign Language (2), Indonesian, Portuguese, Hungarian, and Korean (one each). They had all completed at least high school education, with 74 and 9 holding further undergraduate or graduate degrees, respectively. They self-rated their English ability at a mean of 8.9 (7–9) in speaking, 8.8 (6–9) in listening, 8.9 (7–9) in reading, and 8.7 (4–9) in writing using 9-point scales (1 = *extremely bad*, 9 = *extremely good*), and reported a mean of 99% of daily language use being in English (50–100%), with 96% of all interactions (40–100%) occurring with other native English speakers, using 0–100% scales (0% = *never*, 100% = *all the time*). Native speakers were allowed to complete either one or both of the two non-overlapping versions of the target materials (see below), with 122 participants completing Version A, 123 participants completing Version B, and a total of 104 responding to both. In total, each of the 100 target items was responded to by a minimum of 122 of native speakers.

The L2 English learners included 155 undergraduate students enrolled in an English for Academic Purposes (EAP) program at an English-medium university in Montreal. These participants were sourced from six intact classes in the upper-intermediate level of this program. As university students, all speakers had taken either TOEFL iBT or IELTS tests, demonstrating at minimum a total score of 85 for TOEFL iBT or 6.5 for IELTS, which was considered sufficient for them to pursue academic degrees. By including only those students from a similar course level, we were able to evaluate a relatively homogenous group of L2 users, in terms of proficiency, focusing on their ability to accurately complete the target sequences. L2 learners (87 female, 68 male) came from a variety of language backgrounds, including Chinese (89), Arabic (26), French (14), Farsi (7), Korean (5), Spanish (3), Greek (3), Turkish (2), Polish, Italian, Azeri, Marathi, Bulgarian, and Hausa (one each). While the learners came from a variety of language backgrounds, linguistic background was not a focus of this study. In fact, variability in learners' linguistic backgrounds was seen as a strength, allowing us to examine how a broad range of learners from multiple linguistic backgrounds respond to L2 FSs. L2 learners, who were on average 21.5 years old (18–37), reported a mean of 8.5 years (1–20) of prior English study. Using the same scales, they estimated their English ability at a mean of 6.9 (3–9) in speaking, 7.6 (4–9) in listening, 7.1 (3–9) in reading, and 6.5 (3–9) in writing, and reported communicating with native English speakers on average 52% of the time daily (10–100%). L2 learners, who were tested using identical but paper-based materials in a timed setting, were also randomly assigned to Versions A or B of the materials, with 77 L2 learners completing Version A and 78 completing Version B. Thus, each of the 100 target items was responded to by at least 77 learners.

Materials

The target materials included 100 frequently occurring four-word sequences from the British National Corpus (BNC). Because frequency of occurrence has been used in previous corpus-driven research as a main index in FS identification, the initial step in selecting the sequences was to generate a list of 300 most frequent four-word structures in the BNC using the on-line interface at phrasesinenglish.org (Fletcher 2011). From this list, 100 target structures were chosen through semi-random sampling without replacement, with the criterion that they represented a wide range of values for four metrics: (a) forward transitional probability; (b) backward transitional probability; (c) mutual information; and (d) frequency. Sequences deemed to be overly context specific or unique to British English were removed from the test materials. Since target sequences were taken from a large collection of British English, and the participants were speakers/learners of North American English, this was a necessary step to limit the potential impact of

Table 1. Descriptive Statistics for 100 Target Sequences

Corpus-derived measure	M	Median	Min	Max
Frequency	1131.06	867.00	657.00	6875.00
Mutual information	13.21	13.04	5.17	23.61
Forward transitional probability	.50	.38	.03	1.00
Backward transitional probability	.67	.80	.05	1.00

differences in language variety between North American English and British English. Table 1 summarizes FS statistics for 100 target sequences.

The target structures included various options in terminal position, such as *of*, *a*, *which*, *hand*, *possible* (which were to be completed by participants) to prevent participants from completing sequences based on highly-frequent responses. The 100 target structures were subsequently organized in two randomly-sampled non-overlapping tests (Versions A and B), composed of 50 target sequences and 10 extra fillers drawn from the original list of 300 sequences, for a total of 60 sequences. In each test, all items were listed in a random order with the first three elements in each sequence printed intact and the last element replaced with a blank (e.g. *turned out to* ___, *as soon as* ___, *a great deal* ___). Sample target materials appear in the Appendix, and the complete list can be obtained by e-mailing the corresponding author.

Procedure

Procedures for each participant group varied slightly due to the medium in which the task was administered. The native speakers, tested in the on-line setting, first read a brief project description and digitally signed the consent form by electing to proceed to the online test materials. They were given a maximum of 20 minutes to complete a language background questionnaire followed by the sequence completion task. Before proceeding to the task, they were instructed, in writing, that they would see three-word fragments, followed by a space to write any one word that comes to mind to complete each phrase, and that such words can be long words (e.g. *picture*, *day*, *go*, *thinking*, *fast*, *break*) or short grammatical words (e.g. *at*, *to*, *from*, *the*, *he*, *they*, *she*, *a*). They were then given examples showing possible completions for four unrelated sequences (e.g. *for a long time*, *in the absence of*). The participants then proceeded to type in the missing words for each sequence, working at their own pace, with all completing the test within the allotted period. After finishing one test version (A or B, which the participants accessed first with an approximately equal frequency), they were invited to complete the other version (B or A, respectively), which 104 of the native speakers in fact elected to do, with no restrictions on how soon the second test version could be completed.

The L2 learners, tested as part of several intact groups of 20–25 students during their ESL classes, followed a similar procedure, except that all responses were collected on paper rather than on-line. The learners first read and signed a consent form, then completed the same questionnaire, followed by one of the test versions, for which they were also given a maximum of 20 minutes. The learners received the same instructions, both orally from the experimenter and in writing in their booklets, and were given the same four examples of sequence completions before starting the task. All learners completed the task within the allotted time. Because the learners were tested after all native speaker data had been collected and because in-person testing, compared to the on-line medium, allowed for more flexibility in study design, the ESL classes were randomly assigned to one of the two test materials (Version A or B), which resulted in nearly equal numbers of learners responding to each test version. Each learner completed only a single version of the test to reduce learner fatigue effects, eliminate missed or incomplete datapoints, and allow for testing to be completed within a reasonable time in a language class.

Corpus-based statistics

Four statistics were derived for each target sequence. First, frequency figures, which served as the basis for the computation of all statistics, were sourced from the BNC using the on-line interface at phrasesinenglish.org (Fletcher 2011). Frequency was recorded as the listed frequency count for each target structure. Second, mutual information (MI) figures were computed using the formula MI = log_2(observed frequency/expected frequency). While observed frequency corresponded to the listed frequency count in the BNC, expected frequency required additional calculations. As a first step, probability figures for each of the component words were calculated by taking the frequency of each word and dividing it by the total number of words in the corpus (Schmitt 2010). These figures were then multiplied to achieve the overall probability of all the component words co-occurring. The resulting probability score for the entire sequence was then multiplied by the total number of words in the corpus to derive an expected frequency count for the entire sequence so that the final ratio could be computed. To illustrate this computation with a simple collocation *prime minister* (total frequency = 9,457; *prime* frequency = 11,959; *minister* frequency = 23,401), the probability of *prime* occurring in the corpus is 0.000123 (11,959/96,986,707) while the probability of *minister* is 0.000241 (23,401/96,986,707). Multiplying the two figures yields a general probability of 2.97512^{E-08}, and multiplying this figure by the total number of words in the corpus yields the expected frequency of occurrence (2.89 times), with the resulting MI index of 11.68 (log_2 [9,457/2.89]). This index is high, suggesting that the collocation appears much more frequently than would be expected by chance alone and that the component words are strongly associated.

Finally, two measures of transitional probability were computed. Backward transitional probability (BTP), the probability of the final three words in a structure appearing with the first word, was calculated using the formula BTP(X|Y) = frequency(XY)/frequency(Y), where the numerator denotes the frequency of the entire four-word sequence, and the denominator represents the frequency of the final three words in the same sequence. For example, *the fact that the* appears 2,500 times in the BNC, with *fact that the* having a frequency of 2,666. Using the above formula, BTP equals 0.94 (2,500/2,666), which is high, suggesting that *fact that the* is likely to be preceded by *the*. Similar to BTP, forward transitional probability (FTP), or the probability of the first three words in a sequence being followed by the fourth word, was calculated using the formula FTP(Y|X) = frequency(XY)/frequency(X), where the numerator denotes the frequency of the entire four-word sequence, and the denominator represents the frequency of the first three words. Using the same example, with *the fact that* appearing 12,987 times, the FTP statistic for *the fact that the* is 0.19 (2,500/12,987), which is low, suggesting that the final *the* is only loosely associated with the first three words in the structure. Since BTP and FTP provide distinct measures of association at different points in structures, both were used to investigate their impact on language users' ability to complete the target sequences.

Data analysis

Two dependent variables were used to evaluate the effectiveness of the four metrics to predict language users' ability to complete the target sequences with the intended word. The first variable, proportion of completion, was a measure of how closely participants' completions matched the terminal word in each target structure in the BNC. In the case of minor spelling errors (i.e. *thee* vs. *the*), the answer was changed to the appropriate form and counted as correct. Other answers which mismatched BNC data were counted as wrong. For each target sequence, proportion of completion was derived by dividing the total number of correct answers by the total number of responses available for that sequence (e.g. 0.14 would correspond to 21 correct completions from 155 participants), separately for native speakers and L2 learners. The second dependent variable, range, was a measure of variability in participants' responses. For each target sequence, range was defined as the total number of unique responses given by at least three participants (e.g. 7 would correspond to 7 different completions to a given sequence in three or more participants' responses), computed separately for the two participant groups.[1] Because frequency-based metrics of formulaic status, computed for each target sequence, were the focus of this study, all statistical comparisons were based on item-based statistics.

Results

Preliminary analyses

Before examining the relationship between four corpus-derived metrics (frequency of occurrence, MI, BTP, FTP) and participants' responses in the sequence completion task, two preliminary analyses were carried out. The first focused on the overall performance of the two groups, which is summarized in Table 2. Compared to L2 learners, native speakers were significantly more likely to complete the sequences with the target word, $t(99) = 5.97$, $p < 0.0001$, d (effect size) = 0.63, with a broader range of responses provided, $t(99) = 5.10$, $p < 0.0001$, $d = 1.03$. This confirmed the expected difference in linguistic experience between the two groups, namely, that native speakers were overall more accurate, with a wider range of possible response options available to them, in providing sequence completions consistent with corpus data.

The second analysis examined possible relationships among the four metrics by computing Pearson correlation coefficients between them for the 100 target sequences (see Table 3). With one exception, all measures were correlated with each other, suggesting that they captured a dimension common to all sequences. However, the strength of these relationships was moderate at best. For instance, frequency of occurrence and transitional probability shared only 4–9% of common variance, which confirmed that the two metrics were largely independent of each other. The overlap in shared variance between MI and transitional probability statistics was greater, yet far from perfect (8–25%), implying that the two metrics captured somewhat different dimensions characterizing the target sequences. Unlike the MI statistic, transitional probability is sensitive to the relative order of elements within a sequence, which likely reflected some unique variance (75–92%) in each measure.

Table 2. Descriptive statistics for participants' performance in the sequence completion task

Dependent variable	Native speakers				L2 learners			
	M	SD	Min	Max	M	SD	Min	Max
Completion rate	0.47	0.33	0.00	0.99	0.36	0.33	0.00	1.00
Range	5.70	3.51	1.00	14.00	4.33	2.32	1.00	10.00

Table 3. Pearson correlations between corpus-derived FS metrics

FS metrics	Frequency	MI	FTP	BTP
Frequency	–			
MI	0.12	–		
FTP	0.30**	0.50**	–	
BTP	0.21*	0.29**	0.52**	–

Note. MI = mutual information, FTP = forward transitional probability, BTP = backward transitional probability. * $p < 0.05$, ** $p < 0.01$ (two-tailed).

Table 4. Results of multiple regression analyses using the three corpus-derived metrics as predictors of proportion of completion and range scores for native speakers

Predicted variable	Predictor	Adjusted R^2	R^2 change	β	t	p
Completion rate	FTP	0.65	0.65	0.81	13.72	0.0001
Range	FTP	0.64	0.64	–0.80	–13.38	0.0001

Note. FTP = forward transitional probability.

Corpus-derived metrics as predictors of sequence completions

To determine the relative contribution of the four corpus-derived metrics to native speakers' performance in the sequence completion task, two separate stepwise multiple regression analyses were carried out, with completion rate and range as criterion variables. In these analyses, summarized in Table 4, the four metrics (frequency of occurrence, MI, BTP, FTP) were used as predictor variables. For completion rate, the regression model yielded a single significant predictor, which was the FTP statistic, accounting for 65% of the total variance. For range, the final model accounted for 64% of the total variance, with all of the variance again linked to FTP. In essence, for native speakers, FTP (i.e. the likelihood that the first three words in each sequence are followed by the final word) appeared to singly predict greater proportion of target sequence completions and lower variability associated with these completions.

A comparable set of regression analyses was computed next, with L2 learners' completion rate and range used as criterion variables. These analyses, summarized in Table 5, yielded similar findings. For completion rate, the model revealed a single predictor, FTP, accounting for 50% of the variance in learners' sequence completions. For range, the final model explained a total of 46% of the variance, with 42% linked to FTP and a further 4% associated with the MI statistic. Although the amount of total variance

Table 5. Results of multiple regression analyses using the three corpus-derived metrics as predictors of proportion of completion and range scores for L2 learners

Predicted variable	Predictor	Adjusted R^2	R^2 change	β	t	p
Completion rate	FTP	0.50	0.50	0.71	9.86	0.0001
Range	FTP	0.42	0.42	–0.65	–8.45	0.0001
	MI	0.46	0.04	–0.25	–2.95	0.004

Note. FTP = forward transitional probability, MI = mutual information.

explained by the L2 models was lower compared to native speaker models, the pattern of findings was similar. For L2 learners, as was the case with native speakers, FTP nearly exclusively predicted greater proportion of target sequence completions and smaller range of responses associated with these completions.

Response consistency

The final analysis focused on participants' response agreement in completing the target sequences to determine if a given participant's behavior was consistent with the other participants in the same group. Based on the results of previous analyses, the assumption was that the target sequences featuring higher FTP should elicit greater internal consistency within each participant group. For this analysis, both the overall percent of agreement as well as Cohen's kappa (κ) as an index of interrater reliability appropriate for nominal data were computed separately for each group. Percentage agreement and Cohen's kappa were calculated for each pair of participants, then averaged to yield a single value (Light 1971). The results of these analyses, shown in Table 6, suggested that response consistency was indeed strongly associated with FTP. According to Landis and Koch's (1977) guidelines for interpreting kappa values, native speakers showed 'moderate' agreement (κ = 0.60) for the 33 target sequences featuring high FTP values (above 0.70), while their agreement for the 67 sequences of low FTP (below 0.70) was 'slight' at best (κ = 0.19). This relationship between response consistency and transitional probability was even more pronounced for the 11 sequences from the top and bottom of the transitional probability range, where native speakers demonstrated substantial agreement (κ = 0.72) and virtually random response patterns (κ = 0.09), respectively. As shown in Table 6, L2 learners' response consistency patterned in a similar manner, with the exception that L2 learners, predictably, demonstrated overall lower consistency and that they were less sensitive, in their response agreement, to sequences from the top range of transitional probability.

Table 6. Participant consistency indexes (percent agreement and Cohen's κ) for completion of target sequences of high and low forward transitional probability (FTP)

Sequences	Native speakers		L2 learners	
	% agreement	Cohen's κ	% agreement	Cohen's κ
FTP ≥ 0.98 (*n* = 11)	85.3	0.72	65.9	0.41
FTP ≥ 0.71 (*n* = 33)	69.5	0.60	58.7	0.46
FTP < 0.70 (*n* = 67)	21.8	0.19	19.6	0.17
FTP < 0.14 (*n* = 11)	11.8	0.09	6.0	0.05

Discussion

Taken together, the findings demonstrate the usefulness of transitional probability for corpus-driven identification of FSs. Forward transitional probability was the sole significant predictor accounting for double-digit proportion of variance in native speakers' and L2 learners' responses. Higher transitional probability values were also linked to greater consistency in terms of the range of responses provided, while lower values corresponded to decreased consistency, again for both native speakers and L2 learners. These findings suggest that transitional probability can be used outside word segmentation research to reveal insights into how words pattern together to form units of meaning, thereby improving the identification of FSs in corpora with the view of using such psychologically valid sequences in language classrooms.

Implications for methodology and theory

In previous corpus-driven research, there has been a tendency to identify structures purely according to their frequency of occurrence, applying the range criterion to ensure that the sequences are not restricted to idiosyncratic tendencies of individual users (e.g. Cortes 2004; Biber and Barbieri 2007). However, the current findings suggested that the use of frequency as a measure of formulaic status, with *a priori* decisions regarding sequence length, often fails to accurately identify FSs. A directional measure of word association, transitional probability in fact emerged as a more accurate metric of FSs, insofar as formulaic status can be estimated in a sequence completion task. Compared to mutual information, which provides a general measure of association strength across the entire sequence, and frequency, which reveals how often a particular structure recurs, transitional probability estimates strength of association at crucial points in the structure, leading to more accurate identification of utterance boundaries. Because transitional probability is sensitive to position-specific, directional information, it also contributes to the detection of more functionally complete sequences that no longer cross semantic and syntactic boundaries, thus reducing the incidence of overlapping and partially repeated structures.

The application of transitional probability to formulaic language research holds potential benefits for those attempting to categorize and describe the presence of FSs in various genres and registers. With the traditional lexical bundle approach often misidentifying FSs and producing lists of overlapping structures that lack functional salience, the standard practice of assigning functions to these sequences becomes a challenge that results in the inconsistent assignment of functional roles across studies, such as, for example, treating *the fact that the* and *the nature of the* as discourse organizing bundles or as serving stance and referential purposes, depending on the particular analysis conducted (e.g. Biber et al. 2004; Cortes 2004). With the lexical bundles identified in many studies crossing syntactic and semantic lines, it is not surprising the consistent assignment of functional roles proves to be a difficult task. By making use of transitional probability in the identification of FSs, we can create lists of structures that are more functionally salient and use these results to better categorize and describe the language present in the corpora under investigation.

In the dataset used for this study, forward transitional probability, compared to backward probability, was better at predicting language users' performance. This is unsurprising because completions focused on the last element in each sequence, which was targeted by forward probability. Although backward probability may not have been particularly helpful here, it will likely be as effective for predicting formulaic status of sequences to be completed with the first word of each sequence deleted. For instance, in the sequence *and there is no*, the low backward probability of 0.06 suggests that *and* is not part of this structure. On the other hand, in *as soon as possible*, high backward probability of 0.99 implies that the structure is indeed a four-word sequence. One remaining issue pertains to establishing a threshold of transitional probability to determine a structure's formulaic status. As yet there are no standards; however, current analyses suggest a possible benchmark of 0.70, which was the transitional probability value associated with the sequences that elicited moderate levels of native speaker agreement in sequence completion (see Table 6).

The current results also provide evidence in support of usage-based models of language (e.g. Barlow and Kemmer 2000) which posit that the structures with higher transitional probabilities likely represent usage events that have come to be used by language users as multiword units. In fact, native speakers showed high consistency in their completions of sequences with high forward transitional probabilities (≥ 0.98), suggesting that these sequences might represent units of meaning which have achieved formulaic status. Conversely, sequences with relatively random responses were linked to low transitional probability (< 0.14), implying that these lack functional salience and therefore do not represent usage events in the minds of users. With native speakers and L2 learners demonstrating similar response patterns, frequency of exposure to the target language and sensitivity to frequency-based, statistical regularities in linguistic input (as indexed through transitional probability) emerge as important variables determining language users' ability to complete target sequences. The finding that L2 learners' responses were predictably less accurate and more constrained in their range than the responses of native speakers, likely reflects L2 learners' less extensive and intensive exposure to English, compared to linguistic experience of native speakers (see Ellis 2012). As active L2 users, especially in the academic domain, the L2 learners were likely still in the process of accruing probability statistics needed for them to enjoy the processing benefits of FSs in comprehension and production (e.g. Pawley and Syder 1983; Tremblay, Derwing, Libben, and Westbury 2011).

Implications for teaching

The use of transitional probability in future corpus-driven research has the potential to produce more pedagogically valuable sequences that possess greater functional salience, resulting in structures that will be better suited for pedagogical purposes since they should be easier for students to understand and eventually use in their own discourse. Because previous research focusing on the teaching of FSs has yielded mixed results (e.g. Boers et al. 2006; Cortes 2006), it would seem that which kinds of FSs are being taught, and how, becomes a crucial factor. Even with appropriate instructional techniques, if teachers and students target misidentified sequences, they are unlikely to achieve success. With FSs representative of the language EAP learners are likely to face in their academic careers failing to appear in any meaningful way in many of the most popular English for Academic Purposes texts (Wood and Appel 2014), the lack of emphasis placed on this aspect of language may partially be due to the fact that these sequences are often misidentified in the literature and are therefore perceived as lacking pedagogical value.

The current research highlights a potential value that FSs hold for teaching and suggests that it might be worth using transitional probability to identify functionally usable FSs in a variety of genres and registers, with the idea of ultimately incorporating them in L2 instructional materials. If this is to take place, specific corpora that focus on the registers and genres most relevant to particular groups of learners (e.g. university-level students) will need to be compiled and used as the source texts for the creation of pedagogically relevant FS lists. This is one area where corpus-driven approaches to the identification of FSs may play an important role since it is often difficult to accurately identify important formulaic sequences present in specific contexts through personal reflection alone. This is especially true in ESP contexts, where the language instructor or materials creator may not be well versed in the specific type of English that needs to be taught. By using the corpus-driven methods described in the present study, researchers can objectively identify formulaic sequences that will hold the most benefit to the language learner in these contexts.

Limitations and conclusion

Several limitations of this study must be addressed in future research. First, although the target corpus used in this study was based on British English, all participants were users of North American English. Although attempts were made to control for this limitation, this mismatch in English dialects may have had an effect on at least some participants' ability to complete certain sequences with the target word. Second, the sequence completion task focused exclusively on frequently occurring four-word sequences. The effectiveness of transitional probability to predict sequence completion rates for structures of different lengths remains to be investigated. Finally, transitional probability as a metric of FSs was evaluated in a sequence completion task which likely requires language users to access metalinguistic knowledge about language, instead of targeting the kinds of frequency and usage-based processing implied by input-driven statistics. Therefore, in future studies, transitional probability must be evaluated in tasks which involve a processing speed component (e.g. timed reading or speaking). Despite these limitations, this research points to an advancement in corpus-based identification of FSs, with the use of transitional probability for creating lists of FSs that are more psychologically valid and salient than those identified using traditional methods. To make the most of corpus-based methods, regardless of the specific metric used, future research needs to apply these methods to a variety of corpora, with the goal of developing practical, pedagogic solutions for helping L2 learners across a range of settings.

Note

1. Range was also operationalized as raw value, corresponding to the total number of response options provided, and more conservatively as the number of response options attested in at least 10 participants' data. In all cases, analyses yielded identical findings.

References

Aslin, R. and E. Newport (2012) Statistical learning: from acquiring specific items to forming general rules. *Current Directions in Psychological Science* 21.3: 170–76.

Bamberg, B. (1983) What makes a text coherent? *College Composition and Communication* 34.4: 417–29.

Barlow, B. and S. Kemmer (2000) Introduction: a usage-based conception of language. In M. Barlow and S. Kemmer (eds.), *Usage based models of language*. Stanford, CA: CSLI Publications. 1–64.

Biber, D. (2009) A corpus-driven approach to formulaic language in English: multiword-patterns in speech and writing. *International Journal of Corpus Linguistics* 14.3: 275–311.

— and F. Barbieri (2007) Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26.3: 263–86.

—, S. Conrad and V. Cortes (2004) If you look at . . . : lexical bundles in university teaching and textbooks. *Applied Linguistics* 25.3: 371–405.

Boers, F., J. Eyckmans, J. Kappel, H. Stengers and M. Demecheleer (2006) Formulaic sequences and perceived oral proficiency: putting a lexical approach to the test. *Language Teaching Research* 10.3: 245–61.

Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly* 34.3: 548–60.

Cortes, V. (2004) Lexical bundles in published and student disciplinary writing: examples from history and biology. *English for Specific Purposes* 23.4: 397–423.

— (2006) Teaching lexical bundles in the disciplines: an example from a writing intensive history class. *Linguistics and Education* 17.3: 391–406.

Ellis, N. C. (2012) Frequency-based accounts of second language acquisition. In S.M. Gass and A. Mackey (eds.), *Routledge handbook of second language acquisition*. New York: Routledge. 193–210.

Erman, B. and B. Warren (2000) The idiom principle and the open-choice principle. *Text* 20.1: 29–62.

Firth, J. (1935) The technique of semantics. *Transactions of the Philological Society* 34: 36–77.

Fletcher, W. (2011) *Phrases in English*. Retrieved 18 December 2014 from phrasesinenglish.org

Hyland, K. (2008) As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes* 27.1: 4–21.

Kuiper, K. (1996) *Smooth talkers*. Mahwah, NJ: Erlbaum.

Landis, J. R. and G. G. Koch (1977) The measurement of observer agreement for categorical data. *Biometrics* 33: 159–74.

Light, R. J. (1971) Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological Bulletin* 76.5: 365–77.

Liu, D. (2012) The most frequently-used multi-word constructions in academic written English: a multi-corpus study. *English for Specific Purposes* 31.1: 25–35.

McCully, G. (1985) Writing quality, coherence, and cohesion. *Research in the Teaching of English* 19: 269–82.

Mirman, D., K. Estes and J. Magnuson (2010) Computational modelling of statistical learning: effects of transitional probability versus frequency and links to word learning. *Infancy* 15.5: 471–86.

Nekrasova, T. (2009) English L1 and L2 speakers' knowledge of lexical bundles. *Language Learning* 59.3: 647–86.

Pawley, A. and F. Syder (1983) Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In J. Richards and R. Schmidt (eds.), *Language and communication*. London: Longman. 191–226.

Peters, A. (1983) *The units of language acquisition*. Cambridge: Cambridge University Press.

Schmitt, N. (2010) *Researching vocabulary: a vocabulary research manual*. New York: Palgrave Macmillan.

— and R. Carter (2004) Formulaic sequences in action: an introduction. In N. Schmitt (ed.), *Formulaic sequences: acquisition, processing and use*. Amsterdam: John Benjamins. 1–22.

Simpson-Vlach, R. and N. Ellis (2010) An academic formulas list: new methods in phraseology research. *Applied Linguistics* 31.4: 487–512.

Tremblay, A., B. Derwing, G. Libben and C. Westbury (2011) Processing advantages of lexical bundles: evidence from self-paced reading and sentence recall tasks. *Language Learning* 61.2: 569–613.

Wood, D. and R. Appel (2014) Multiword constructions in first year business and engineering university textbooks and EAP textbooks. *Journal of English for Academic Purposes* 15.1: 1–13.

email: je.appel@gmail.com

[Received 3 December 2014]

Appendix

Sample target sequences, with corresponding corpus-based statistics

Sequence	Frequency	MI	FTP	BTP
the extent to which	1746.0	14.76	1.00	0.98
it would have been	1382.0	15.44	0.52	0.20
there is no doubt	721.0	19.53	0.06	0.99
in the wake of	729.0	12.79	0.99	0.99
in the house of	898.0	8.79	0.29	0.22
it may be that	758.0	12.78	0.14	0.89
with the help of	844.0	10.72	0.99	0.72
in the sense that	830.0	11.36	0.57	0.91
but it is not	790.0	11.77	0.13	0.08
on the other hand	5284.0	17.51	0.71	0.98

Note. MI = mutual information, FTP = forward transitional probability, BTP = backward transitional probability.