GLOTTODIDACTICA XLV/2 (2018)
ADAM MICKIEWICZ UNIVERSITY PRESS POZNAŃ
DOI: 10.14746/gl.2018.45.2.07

ROMUALD GOZDAWA-GOŁĘBIOWSKI

Uniwersytet Warszawski
r.gozdawa@uw.edu.pl
ORCID: 0000-0002-6965-8025

MARCIN OPACKI

Uniwersytet Warszawski
marcin.opacki@wn.uw.edu.pl
ORCID: 0000-0003-3122-3568

Recurrent strings in corpus-based pedagogical research: A reappraisal of the field

ABSTRACT. Formulaic competence is a hotly debated issue in teaching circles, not only because of its role in L2 communication but also due to the inherent complexity of the identification criteria for formulaic strings. While the mixed approach, combining meaning-based and corpus-based identification measures, remains a natural solution, the subjective character of the criteria, together with the required involvement of native experts, diminishes its attractiveness for every-day pedagogical purposes. We would like to explore the potential of "corpus-only" identification tools. Specifically, our objective is to show that meaningless n-grams (*of the*, *in a*, etc.) generated by frequency searches contain useful pedagogical data, and that, coupled with MI scores frequency-based measures accurately characterize learners' formulaic competence. Because of the relative simplicity of the identification procedure and free availability of corpus tools, frequency-based and distribution-based measures may become an important new pedagogical tool at the disposal of language teachers.

KEYWORDS: formulaicity, recurrent strings, corpus linguistics, frequency, distribution.

1. OVERVIEW OF THE PAPER

Given the communicative focus of present-day FL pedagogy, prioritising task completion and dethroning morpho-syntactic accuracy, formulaic competence has come under the spotlight. The recognition of the role of chunks

in language production and comprehension (cf. Wray 2002) has in turn provoked a heated debate about the most efficient tools for identifying formulaic strings in any stretch of text. Whatever tool is chosen, however, care is usually taken to exclude from the set of possible formulae one type of word sequences which rank highest on any frequency-based or distribution-based list: combinations of grammar words such as *in a*, *of the*, etc. In this paper, we would like to restore these unwanted sequences to grace, arguing for their usefulness for pedagogical purposes. Also, we want to run a corpus-based analysis (cf. Opacki 2017) of B2 and C1 essays by Polish learners of English with respect to the distribution of formulaic language to verify how the conclusions based on the strength of associations, without any human intervention, compare to the findings reported in the relevant literature.

2. THE THREE CHALLENGES OF FORMULAICITY

Formulaic language is a confusing concept at least on three levels: what it is, how it is conceptualised and how it is measured. Hence we see three major problem areas that research into formulaicity needs to address:

a) **pedagogical challenge** – the very nature of formulaicity makes it a hard area to teach: it is unpredictable in character, lexically based rather than rule-based, culture-dependent, conventionalised, genre-dependent, prone to L1 interference (negative transfer),
b) **ontological challenge** – the elusive nature of the concept and the broad range of entities it tries to embrace have been responsible for a whole array of definitions, models and functions which formulaic language is said to perform,
c) **methodological challenge** – the proposed methods of identifying formulae, varying in scope and character, are difficult to implement in actual teaching practice and/or require formal linguistic knowledge and a native feel for the language.

2.1. Pedagogical challenge: Inconclusive experimental data

There have been numerous attempts to improve the rate of "formulaic intake" in the foreign language classrooms, with mixed results, as suggested by the overview of the findings below.

Forsberg (2010) argues that the increase in the number of formulaic expressions (her *conventionalized sequences*) is a fair indicator of second language development (essentially following Yorio 1989), as predicted by

a frequency-based SLA theory. In fact, all findings reporting a growth in the recognition and/or use of formulae by L2ers relate that growth to the duration and intensity of exposure to the target language. For example, Jones and Haywood (2004) showed that overt tuition raised the retention level of formulaic expressions in a group of EAP (English for Academic Purposes) students. Kazemi, Katiraei and Rasekh (2014) observed a similar trend in a group of Iranian EFL students, Nasiri and Khorshidi examined the role of ZPD (zone of proximal development) sensitive feedback in this process. Peters and Pauwels (2015) show how overt (explicit) vocabulary teaching positively influences the use of formulaic sequences in learners' written output. Szudarski and Carter (2016) use input flood and input enhancement to promote (passive) collocational knowledge of V–N and A–N sequences.

On the other hand, the difficulties in the productive use of formulae have been noted in Nesselhauf (2003), Granger (1998) and in Gilquin and Paquot (2008), who report the persistence of non-native patterns in the formulaic language of L2 learners. This non-formulaic bias, first noted by Pawley and Syder (1983), is attributed in Skehan and Foster (2001) to learners' preference for rule-based sequences, which are a substitute for native intuitions. The notorious underappreciation of formulae by adult L2ers is discussed in Weinert (1995) and Wray (2002). The fact that even a prolonged stay in the native-language environment does not bring about the expected gains in formulaicity is related in Hulstijn and Marchena (1989) to risk avoidance and the adoption of a "play-it-safe" strategy (cf. also Siyanova & Schmitt 2007). Even advanced users, with a large range of appropriate collocations at their disposal are shown in Siyanova and Schmitt (2008) to lack native-like intuitions and fluency.

2.2. Ontological challenge: Defining formulaicity

Perhaps the best illustration of the difficulties involved in defining and indentifying formulaic sequences comes from Wray (2008): "[It is] rather like trying to find black cats in a dark room. You know they're there but you can't pick them out from everything else". Some confusion is simply unavoidable.

There are essentially four major approaches to formulaic language: phraseological, psychological, sociolinguistic and corpus-based (cf. for example Wood 2015; Forsberg 2010; Durrant & Mathews-Aydınlı 2011). The psychological angle, culminating in Wray's (2002: 9) famous observation that a formula is a sequence of elements which are "stored and retrieved whole from memory (…) rather than being subject to generation or analysis by the language grammar" and reiterated in Wray's (2008) definition of *a morpheme equivalent unit* ("a unit processed as a morpheme without recourse to any form-meaning

matching of the subparts") is often sidestepped (without questioning its validity) in pedagogical research, as too vast and difficult to verify and operationalize (Schmitt, Dörnyei, Adolphs & Durow 2004; Forsberg 2010).

Pedagogically oriented research has also suggested a sociolinguistic dimension to defining and identifying formulaic language, as in Nattinger and DeCarrico's (1992) lexical phrases, Schmitt and Carter's (2004) list of purposes which formulaic language serves in communication, Pawley's (2007) formulae with pragmatic functions or Wray's (2008) concept of the manipulative (non-linguistic) function of formulaic language. This can be seen perhaps most clearly in Durrant and Mathews-Aydınlı (2011), who adopt a 'function-first' approach to identifying formulae: a relevant corpus is annotated for communicative functions and then formulae are identified "as the recurrent patterns associated with each function." While this is a promising direction to take and it does lead to interesting generalizations (cf. Durrant & Mathews-Aydınlı's pedagogically oriented definition of formulaicity), it requires a multi-step reiterative preparatory analysis and is unsuitable for immediate pedagogical implementation.

As for the phraseological approach to defining formulaicity (Howarth 1998; Nesselhauf 2003; 2005), it is built around the concepts of fixedness, substitution and compositionality and thus requires a complex checking procedure, with a major role of informed native intuitions. For example, Schmitt and Martinez (2012) in an attempt to establish their Phrasal Expressions List (a commendable project and one that certainly deserves a continuation) require native judges to determine the degree of "morpheme equivalence" of every potential candidate for a phrasal expression as well as their semantic vs. deceptive transparency. In a similar project, Simpson-Vlach and Ellis (2010) crucially relied on native judgments for defining and identifying academic formulae and their subtypes.

Each of the trends delineated above has its own merits and should not be discouraged. The alternative we are going to explore here is that a lot of accurate and pedagogically useful information about the role of fixed expressions in L2 English may come from corpus findings, which define formulae in terms of their frequency and distribution across a corpus or a number of corpora. We address this issue in the next subsection.

2.3. Methodological challenge: Identifying formulae in a text

There are presumably as many methods of identifying formulae as there are ways to define them (cf. Section 2.2). Pawley (2007) proposes seven criteria, Wray (2002) and Wray and Namba (2003) list eleven criteria which are

supposed to adequately capture the spirit of formulaicity. This is a comprehensive checklist, as noted in Wood (2015: 26–27), but based on the recognition of *gradience* of formulaic expressions (i.e. some of them may be more formulaic than others), with experts passing value judgments on a Likert scale of 1 to 5. The criteria proposed by Wray and Namba (2003) will, for the most part, illustrate the phraseological method of identifying formulaic sequences, because it involves manual identification. For the purposes of this paper, we will refer to this method as a meaning-based collocational measure, since a formulaic sequence is in essence a phraseologism and we find the concept of "a phraseological method of identifying phraseologisms" to be semantically redundant.

An alternative to the meaning-based procedure is the statistical method (cf. Granger & Pacquot 2008; Gablasova, Brezina & McEnery 2017; Wood 2015; Nesselhauf 2005; Paquot & Granger 2012). In its simplest form, it is sensitive only to quantitative evidence (word frequencies) and it yields recurrent sequences (De Cock 1998), i.e. the most common 2-, 3- or 4-word sequences or surface co-occurrences of word forms (n-grams) in a given corpus.

We believe that there is a downside and an upside to the frequency-only approach and we will examine both in the next subsection. For the time being let us merely note that many scholars (e.g. Siyanova & Schmitt 2008; Wood 2015) supplement the frequency criterion with "distributional" association measures (AMs) which "combine information about frequency with other collocational properties that can be expressed mathematically" (Gablasova et al. 2017), such as MI (Mutual Information score), assessing the strength of the links between immediately adjacent pairs of words in any given corpus.

2.4. For and against corpus-based collocational measurement tools

Since it is our intention in the paper to demonstrate the usefulness of the distribution-based and frequency-based tools for identifying word combinations, let us look at some of the problems which the statistical approach faces.

Bretaña and Bertrán (2008) voice the most commonly raised objection that many of the automatically extracted sequences are "not interesting from the point of view of phraseology". The sentiment is reiterated in Wood (2015: 21), who observes that "additional steps are also required to eliminate meaningless combinations of words for functional analyses of formulaic language". The objection makes perfect sense if a refined analysis will follow the preliminary categorization, in which case the frequency-based count is merely a necessary first step. Alternatively, as in the case of Forsberg's (2010)

research on conventionalized sequences, decisions have to be taken individually by a human researcher for every word cluster and the applicability of statistical tools is minimal at best. However, we need to explore the possibility that those "phraseologically uninteresting" and "meaningless" combinations of words may be a source of pedagogical insights or important L2 generalizations and as such should not be *a priori* shrugged off as irrelevant.

Another frequently raised objection is that statistical methods are typically used only on large corpora such as the BNC (British National Corpus) (Forsberg 2010; Wood 2015) and a statistical analysis carried out on a small-sized group of learners or on a micro corpus is bound to produce incorrect results. Let us note, in defence of the corpus-based approach, that statistical tools are permissible if the size of the sample is equal to the size of the population, the real question being how far we want to take our generalizations, based on group findings.

Wray points out that frequency counts may distort the picture by disallowing formulae which need to be recognized, for example those that would be infrequent in any general English corpus, yet are known to perform an important socio-cultural role in an English speaking community. The problem may be easily overcome by supplementing frequency counts with other associative measures, such as MI scores, as they are sensitive to collocational strength of a pair of words, rather than their frequency. The same point has been made in Wood (2015: 21).

3. A CHANGE OF PERSPECTIVE?

We do not question the need to introduce fine-tuned definitions of formulaic expressions. Nor do we object to introducing a range of methods to identify and categorize these expressions. Note that the essential problem encountered in the approaches overviewed here is that extra mechanisms are needed to supplement (or replace) the corpus-based measures because the strings of words generated even by the refined or combined statistical tools significantly overgeneralize, i.e. allow chunks which do not fall under any of the categories of formulaic language that the teacher/researcher is ready to accept. We want to make two claims in this context:

a) the "unwanted" strings, which the corpus-based tools freely generate, are vital for grammar awareness purposes, despite their failure to satisfy even the broadest definition of formulaic language; it may be the case that an alternative label should be sought to stress their structural (system-building properties). We will refer to them as morpho-syntactic cues (*MOSC*). Alternatively, a more encompassing definition of

a formula, or a fixed cluster, may be considered, which would incorporate both "formulae proper" and MOSCs: a fixed cluster (FC) is a sequence of words – continuous or discontinuous, with a statistically significant strength of association, as measured by the Mutual Information score and/or the *t*-score. This avoids the trap of frequency-based definitions (by determining the strength of mutual associations) but it is open to criticism on semantic grounds –many of the emerging word sequences will be "uninteresting" from the phraseological perspective, i.e. not formulae in any accepted sense of the word. We believe the criticism to be too harsh (cf. the next section);

b) the corpus-based association measures (cf. Gablasova et al. 2017), if properly used, provide a wealth of information about the formulaic competence of our learners, identify their weak and strong points, perform numerous tasks traditionally delegated to meaning-based approaches, help distinguish between competence levels (e.g. B2 and C1, using CEFR descriptors), enable meaningful comparisons with native reference corpora.

In the empirical part of this paper, we will put these claims to the test, using for illustrative purposes a corpus of writing samples from a total of 92 participants – students of the Institute of English Studies at the University of Warsaw (Poland) and a reference corpus of native-speaker essays.

4. EXPERIMENTAL DATA

We accumulated a corpus of writing samples from a total of $N = 92$ participants from four groups of $N = 23$ participants each. The participants were matched for age ($M = 20$). The first group ($N = 23$) consisted of students who took a preliminary one-semester course in academic writing at the Institute of English Studies of the University of Warsaw. The overall language proficiency of this group could be assessed as borderline B2/C1. The second group ($N = 23$) consisted of students from the same institution (IES UW), but with certified C1 proficiency (post C1 exam), attending a post-graduate CLIL course in writing for research. The third group ($N = 23$) consisted of applicants to universities in the United States who sought to improve the composition of their admissions essays (a requirement at selected institutions) with the aid of an online writing forum moderated by professional English teachers. The fourth group ($N = 23$) consisted of B2-level certification exam papers from students representing various departments of the University of Warsaw.

Each group wrote the same type of composition. While topics were not identical for all groups, their range was similar in all groups in that they

were either opinion essays that expressed views on a relatable concept or utility compositions (i.e. applications, letters of complaint, requests, etc.). The form of all compositions was consistent throughout, namely an essay of 350–500 words (tokens) consisting of an introduction ("intro"), main body (of no less than one paragraph), and a concluding paragraph (or "outro"). Considering the aforesaid conditions, we deemed cross-group comparisons viable. All corpus documents were encoded in the .xml (UTF-8) format, which included basic metadata, such as topic, author, date composed, score (where applicable), but no in-text linguistic annotation (this is due to the fact that we aimed to emulate the limited resources of a typical classroom context). All corpus-specific analyses were performed using the readily available freeware annotator AntConc (Anthony 2018).

4.1. Comparing the tools

For our first analysis, we sought to compare n-gram based queries and their MI-based equivalents with respect to the number of returned search hits when analysing one of our learner sub-corpora, the B2-level essays. While both query types might be, given the right circumstances, useful from a pedagogical standpoint, it was our intention and hope to pinpoint which of these methods would yield readily interpretable, accurate results while retaining the much desired methodological simplicity, which may be appreciated by the busy, yet dedicated, practitioner. The minimum frequency and minimum range for both searches was set to ten. Additionally, the search span, applicable to the MI, was set to five both on the left and the right of the queried word and the n-gram size range was set to 2 (bi-gram)– 5 (pentagram).

Given the fact that n-gram measures represent linear clusters and that MI scores stand for discrete relationships between tokens, we expected the latter to return more hits than the former. This expectation was met, as revealed by the raw data presented in Table 1 below.

To compare these sets of raw data, we first established their normality status using the Shapiro-Wilk test with a false positive tolerance of $a = .05$. The normal distribution condition failed for both data sets ($W = 75$, $p < .001$; $W = 99$, $p = .02$). Accordingly, we opted to conduct a non-parametric comparison in the form of the Mann-Whitney *U*-Test. The test revealed a significant difference between conditions, $U = 1$, $r = .84$, $p < .001$. We can thus observe that MI-based measures return more collocates per each word queried than their n-gram-based equivalents, as shown in the chart below.

Table 1. Comparing n-gram and MI scores

WORD	COLS FOUND USING MI (L/R SPAN 5, MF/MR 10)	COLS FOUND USING N-GRAMS (2-15, MF, MR 10)	DIFF
Best	54	23	31
Education	47	3	44
Form	28	17	11
Good	48	2	46
Have	160	35	125
Last	42	22	20
Learn	66	10	56
Life	56	2	54
Like	56	7	49
New	75	7	68
Opinion	41	18	23
People	157	15	142
See	54	7	47
Service	49	9	40
Stay	43	14	29
Time	74	9	65
Travelling	102	27	75
Want	44	3	41
Way	65	19	46
World	76	15	61

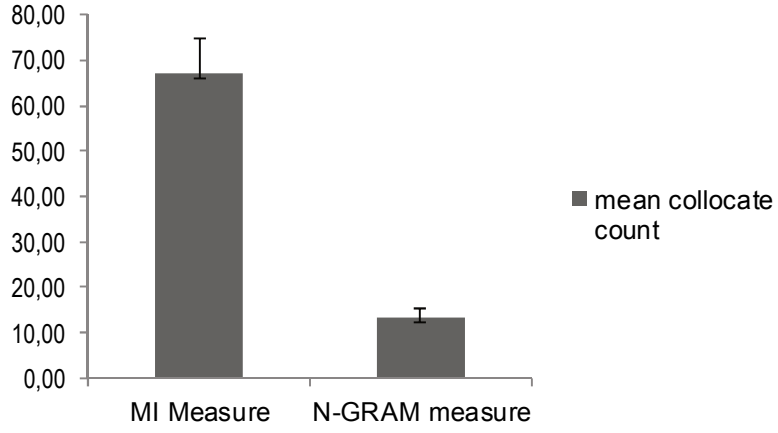

Figure 1. Collocation counts in two query modes

Thus, there is a basis to conclude that MI measures return considerably more potential collocations when compared with n-gram-based measures. Indeed, upon a qualitative investigation of the encountered discrepancy, it becomes clear that several items of potential pedagogical interest were missed by the n-gram query. Some of these omitted items are listed below.

- *I [am/was] (extremely) pleased with my last stay at your hotel*
- *I [am/was] (extremely) disappointed with my last stay at your hotel*
- *meet young people*
- *spend free time*
- *[adj.] social life*
- *[keep/stay/be] in touch with [n.]*
- *last but not least*
- *["what" consistently used as a complementizer, e.g. *"Something what is not very interesting"]*

While the phraseological or collocational status of each item listed above can be debated, all of these clusters are extremely informative with respect to the learner group investigated, revealing a great deal about the preferred patterns of use. Of particular note is the last item (missed by the n-gram measure), which indicates that the learners treat the interrogative/relative *what* as a universal complementizer (i.e. *what*, in the "collective interlanguage" of the learners, is equivalent to *that*). To the practitioner, and in particular the teacher or educational analyst, the discovery of this pattern has the potential to help with pedagogical intervention in the classroom, showcasing the usefulness of the MI measure as a powerful tool for diagnostic and assessment purposes. Since this result was returned only by the MI measure and not the n-gram measure, we suggest MI, and its derivatives, such as the *t*-score, for pedagogical purposes despite their superficial, yet nonetheless occasionally off-putting, complexity. This intricacy, admittedly relative, should no longer result in the avoidance of corpus tools by in-service teachers, as user-friendly software, such as AntConc (for the latest version see Anthony 2018) is widely (and freely) available.

4.2. Fixed expressions as a measure of stereotypicality in writing

For our next analysis, we compiled a list of widely accepted formulaic expressions extracted from the British National Corpus that were assessed as "stereotypical" by a trained linguist. By "stereotypical", we mean that these are phrases used as crutches when building sentences. We then proceeded to compare the frequency and range of their occurrence across our four subcorpora. Through this, we hoped to investigate whether there are any differ-

ences in how these formulae, commonly taught in English language classrooms, are distributed across various proficiency levels. The raw data from our analyses is presented in Table 2 below.

Table 2. Stereotypical formulae across competence levels

Fixed expression/Formula	B2+/C1 Writing practice	B2 exam	Native Uni apps.	C1Writing for research course
time (e.g. at that time, at that particular point in time, at the same time, for the time being)	32	3	41	2
what is more	36	92	0	6
first of all	17	134	0	1
all in all	6	14	0	0
for the time being	5	0	0	0
then again	1	0	0	0
on the other hand	31	64	2	2
when it comes to	22	5	3	6
[X-subj] would like to	33	129	2	1
last but not least	13	14	0	0
in this [essay|case|area|instance|field]	71	59	9	18
as a result of [X]	12	1	1	0
one of the (adv) [adj] [X] [prep] (e.g. one of the most important Polish composers, one of the oldest and biggest religious institutions in the world, one of the most bizarre and destructive phenomena in the universe, one of the nicest professors, etc.)	67	43	18	9
thank you (very) (much|kindly) for your [X]	15	3	1	0
the fact that [X]	60	24	1	13
there is no [X]	37	47	2	7
as far as [X] [is concerned]	16	4	0	6
look forward to (e.g. I look forward, she looks forward, he is looking forward to, she would be looking forward to, etc.)	15	31	2	0
take advantage of (incl. active and passive voice, e.g. can take advantage of [X], [X] has been taken advantage of, etc.)	11	6	0	0
a way to [X-INFP] (e.g. a way to visualize it, a way to make use of it, a way to escape those limitations, a way to minimize losses, etc.)	11	3	4	0
more and more (e.g. we discover more and more about our universe)	11	44	2	0

cont. tab. 2

Fixed expression/Formula	B2+/C1 Writing practice	B2 exam	Native Uni apps.	C1Writing for research course
in order to (e.g. adults try to stop playing in order to give attention to more important things, being a scholar means constantly analyzing those changing elements in order to extrapolate adequate conclusions that can be then followed as universal principals of certain domain.)	55	14	7	8
on (the) one hand	3	6	2	0

We used recursive query syntax to locate the expressions using the MI-based method and using the same pre-sets as in our previous analysis, i.e. a minimum range of and frequency of ten and a leftward and rightward span of five from the collocation root word. A random sampling script was used to select 300 documents from each sub-corpus.

Once again, all samples were tested for normality using the Shapiro-Wilk method and, as is typical of independent corpora, a normal distribution could not be ensured, $W = 0.86993379$, $p < .001$ (B2+/C1 Group), $W = 0.776780006$, $p < .001$ (B2 exam Group), $W = 0.498861671$, $p < .001$ (Natives), $W = 0.744794839$, $p < .001$ (C1 Writing for research group). We performed a non-parametric test in the form of the Kruskal-Wallis *H*-test (Bonferroni correction applied).

The test revealed a significant difference among conditions, $H(3) = 36.41$, $p < .001$. Due to the significant difference, a follow-up analysis was called for. At this juncture, we opted for the Nemenyi Test, which yielded the following contrasts: no significant difference between the B2 (cert) condition and the C1 (writing practice) condition, $q = 1.15$; no significant difference between the Natives condition and C1 (CLIL Linguistics) condition, $q = 0.17$; a significant difference between the C1 (WP) condition and the Natives conditions, $q = 6.46$; a significant difference between the B2 condition and the Natives condition, $q = 5.31$; a significant difference between the C1 (WP) condition and C1 (CLIL LING) condition, $q = 6.64$; a significant difference between the B2 condition and C1 (CLIL LINGUISTICS) condition, $q = 5.48$. Figure 2 captures these results.

There is a notable difference between the two lower proficiency levels (B2+/C1 and B2). However, a more pronounced dissimilarity exists between the C1 CLIL learners on the one hand and the two lower level groups combined. Of particular interest is the emergent fact that there is no meaningful difference between C1 CLIL learners and native speakers. We take this as

evidence that as learners advance, they cease to rely on stereotypical expressions learned by rote in a classroom environment and begin to exhibit notable traits of the same creativity that is typical of native speakers in contexts where linguistic sophistication is expected and encouraged. We thus draw a general conclusion that the degree of stereotypical language use is inversely proportional to proficiency.

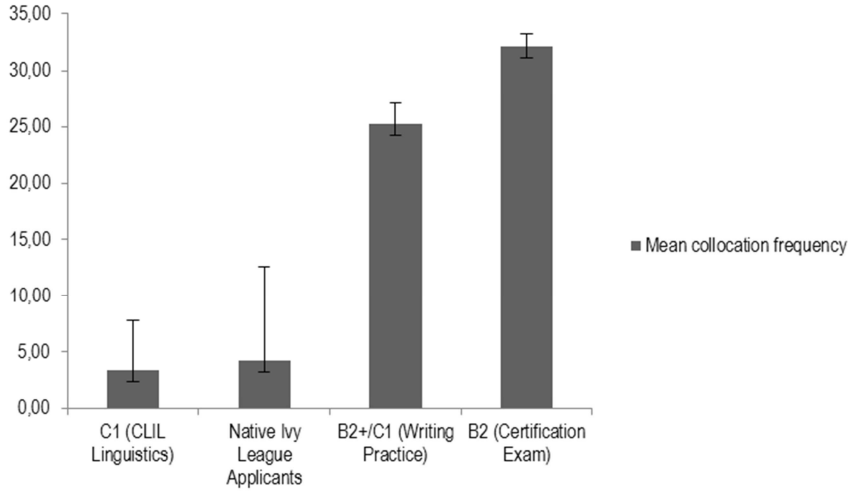

Figure 2. Mean collocation frequency across the four sub-corpora, showcasing the stereotypical use of fixed expressions at lower levels of proficiency

4.3. Using n-grams to assess fixed expressions

Given the outcomes of the comparative MI and n-gram queries, it is tempting to conclude that n-grams are the inferior measure. We would like to advise against this conclusion. Indeed, the measures afford dissimilar results, but one must also bear in mind that their applications are complementary in nature, rather than mutually exclusive. The n-gram tool certainly has the advantage of being readily implementable without much preparation or training, making it well-suited for classroom purposes. The correspondence between n-grams and the MI score is roughly equivalent to that between a longitudinal study and its cross-sectional counterpart: one involves more resources while the other is quick and simple. There is a time and place for either or both.

We have come to appreciate simple n-gram analyses for two main reasons: they help to resolve the methodological challenge of identifying

formulae (Section 2.3.) and they draw attention to morpho-syntactic cues (MOSC), which fail to satisfy any meaning-based definition of a fixed expression. While the pedagogical relevance of commonly recognized formulae is seldom called into question, what about the relevance of all those numerous chunks that would not be deemed formulaic by native speakers? Typically, these items are brushed aside as artefacts. And yet, it is vital to acknowledge their importance in utterance building and in raising learners' awareness of the L2 system.

This importance manifests itself in the fact that only lower-level learners (in CEFR terms) rely on stereotypical chunks (i.e. entire phrases), though both high and low proficiency speakers use chunks. Advanced learners seem to rely on "sentence stems" which are better suited for all manner of creative expression, but ill-suited when it comes to accommodating the lexico-grammatical needs of less advanced learners. We believe it to be a major pedagogical finding, since reliance on sentence stems rather than formulae seems to be a native-speaker trait. This should become apparent upon even a cursory inspection of the token clusters found in our native speaker group (Table 3 below).

Table 3. Selected items from the native n-gram list showcasing sentence stems rather than complete formulae as the predominant structures used by native speakers

at the same	in my opinion	one of the most	go to the
because it is	in the future	that there is	have to be
because of the	in the past	the fact that	I look forward to
but it is	in the world	this is the	I would like to
due to	in this	when it	if you want to
it is possible	seems to	in my life	is the most
it is important	thank you for	is the best	

While some of the items can be considered formulaic (in particular hedges such as *in my opinion*), expressions such as *is the most*, *it is possible*, *it is important*, *because it i*s and *have to be* are the glue that binds lexical items and formulae. These are themselves unlikely to be classified as formulae in a semantic or phraseological sense; still, we argue that they reveal a great deal of information about the proficiency level of the subjects investigated. Simply put, advanced users will use more of them, and rely less on what is typically understood as formulae. In this way, learner progression or advancement from lower to higher proficiency levels can be regarded as

a matter of shifting from the replicative and stereotypical to the productive and expressive.

Finally, let us note that even at lower levels of competence (A1–B1 plus) the morphosyntactic clues, generated by n-gram searches can be a source of didactic inspirations. With reference to our data the first ten bi-grams do not correspond on a one-to-one basis to any sequence of Polish words. The first three of those are: "of the", "in the" and "it is". Let us briefly focus on the first one. It really is a meaningful sequence, as it denotes a relation of possession with some definite NP, like a pattern generator with an open slot. Getting learners to recognise the impact of the analytical genitive (*of*) and the lexically overt marker of definiteness (*the*) on sentence well-formedness is a huge pedagogical challenge, since native Polish habits suggest replacing *of* with the inflectional ending *-s* and articles are absent from most Slavonic languages (for more discussion cf. Opacki & Gozdawa-Gołębiowski 2017). Monitoring her group's progress with an occasional corpus check of the distribution of the analytic vs. periphrastic genitive lets a teacher make informed decisions about the content of her language courses and possible remedial measures.

5. CONCLUSIONS

In this paper, we tried to show how frequency-based (n-grams) and distribution-based (MI scores) corpus tools may be applied in every-day pedagogical practice to yield important information about the role and patterns of use of recurrent expressions in writing tasks of Polish advanced users of English. While we are not implying that native speaker judgments should be neglected in phraseological studies, we suggest alternative solutions in cases where on-going formulaic assessment is necessary in the absence of native informants. A sample corpus-based analysis we performed for the purposes of this paper returned important results about the linguistic profile of B2 and C1 groups of Polish learners, showed specific problem areas (e.g. the overgeneralization of relative "what") to be addressed in the classroom and pointed to a major drop in the use of stereotypical fixed patterns in the written output of more advanced users in comparison with B2 users. A similar drop was observed in the native English reference corpus. The role of n-gram analyses for raising grammatical awareness and identifying preferred multiword expressions was also explored. With n-gram analyses being simple to run, this suggests new directions in the present-day language-teaching methodology.

REFERENCES

Anthony, L. (2018). AntConc (Version 3.2.1) [Computer Software]. Tokyo, Japan: Waseda University. http://www.antlab.sci.waseda.ac.jp.

Bretaña, J.-M. P. / Bertrán, A. P. (2008). Combined statistical and grammatical criteria for the retrieval of phraseological units in an electronic corpus. In: S. Granger / F. Meunier (eds.), *Phraseology: An interdisciplinary perspective* (pp. 391–406). Amsterdam: John Benjamins.

De Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics, 3*, 59–80. https://doi.org/10.1075/ijcl.3.1.04dec.

Durrant, P. / Mathews-Aydınlı, J. (2011). A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes, 30* (1), 58-72.

Forsberg, F. (2010). Using conventional sequences in L2 French. *Iral-international Review of Applied Linguistics in Language Teaching – IRAL-INT REV APPL LINGUIST, 48*, 25–51. doi.org/10.1515/iral.2010.002.

Gablasova, D. / Brezina, V. / McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing and interpreting the evidence. *Language Learning, 67*, 155–179. https://doi.org/10.1111/lang.12225.

Gilquin, G. / Paquot. M. (2008). Too chatty. Learner academic writing and register variation. *English Text Construction*, *1* (1), 41 – 61. doi.org/10.1075/etc.1.1.05gil.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: collocations and formulae. In: A. Cowi (ed.), *Phraseology: Theory, analysis and applications* (pp. 145–160). Oxford: Oxford University Press.

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

Hulstijn, J. H. /Marchena, E. (1989). Avoidance: grammatical or semantic causes? *Studies in Second Language Acquisition*, *11* (03), 241–255. doi.org/10.1017/S0272263100008123.

Jones, M. / Haywood, S. (2004). Facilitating the acquisition of formulaic sequences. In: N. Schmitt (ed.), *Formulaic sequences* (pp. 269–292). Amsterdam/Philadelphia: John Benjamins.

Kazemi, M. / Katiraei, S. / Rasekh, A. E. (2014). The impact of teaching lexical bundles on improving Iranian EFL students' writing skill. *Procedia – Social and Behavioral Sciences*, *98*, 864–869. doi.org/10.1016/j.sbspro.2014.03.493.

Martinez, R. / Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics*, *33* (3), 299–320. doi.org/10.1093/applin/ams010.

Nasiri, M. / Khorshidi, S. (2015). Dynamic assessment of formulaic sequences in Iranian EFL learners' writing. *International Journal of Language and Applied Linguistics*, *1*, 26–32.

Nattinger, J. R. / DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, *24* (2), 223–242. doi.org/10.1093/applin/24.2.223.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Philadelphia, PA: John Benjamins.

Opacki, M. (2017). *Reconsidering early bilingualism: A corpus-based study of Polish migrant children in the United Kingdom*. Frankfurt am Main: Peter Lang.

Opacki, M. / Gozdawa-Gołębiowski, R. (2017). Towards a distribution-based corpus analysis of transfer-susceptible NP modifiers. A case of Polish advanced users of L2 English. *Konin Language Studies*, *5* (1), 9–35.

Paquot, M. / Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, *32*, 130–149. doi.org/10.1017/S0267190512000098.

Pawley, A. (2007). Developments in the study of formulaic language since 1970: A personal view. *Phraseology and Culture in English*, 3–48. doi.org/10.1515/9783110197860.3.

Pawley, A. / Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In: J. C. Richards / R. W. Schmidt (eds.), *Language and communication* (pp. 191–225). London: Longman.

Peters, E. / Pauwels, P. (2015). Learning academic formulaic sequences. *Journal of English for Academic Purposes*, *20*, 28–39. https://doi.org/10.1016/j.jeap.2015.04.002.

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

Schmitt, N. / Dörnyei, Z. / Adolphs, S. / Durow, V. (2004). Knowledge and acquisition of formulaic sequences: A longitudinal study. In: N. Schmitt (ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 55–86). Amsterdam: John Benjamins.

Simpson-Vlach, R. / Ellis, N. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31*, 487–512.

Siyanova, A. / Schmitt, N. (2007). Native and nonnative use of multi-word vs. one-word verbs. *International Review of Applied Linguistics*, 45, 119–139. DOI: 10.1515/IRAL.2007.005.

Siyanova, A. / Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study Perspective. *Canadian Modern Language Review*, *64* (3), 429–458. https://doi.org/10.3138/cmlr.64.3.429.

Skehan, P. /Foster, P. (2001). Cognition and tasks. In: P. Robinson (ed.), *Cognition and second language learning* (pp. 183–205). New York: Cambridge University Press.

Szudarski, P. / Carter, R. (2016). The role of input flood and input enhancement in EFL learners' acquisition of collocations: L2 input types and acquisition of collocations. *International Journal of Applied Linguistics*, *26* (2), 245–265. https://doi.org/10.1111/ijal.12092.

Weinert, R. (1995). The role of formulaic language in second language acquisition: A review. *Applied Linguistics*, *16* (2), 180–205. doi.org/10.1093/applin/16.2.180.

Wood, D. (2015). *Fundamentals of formulaic language*. London, New York, New Delhi: Bloomsbury Academic.

Wray, A. (2005). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford, New York: Oxford University Press.

Wray, A. / Namba, K. (2003). Use of formulaic language by a Japanese-English bilingual child: A practical approach to data analysis. *Japanese Journal for Multilingualism and Multiculturalism*, *9* (1), 24–51.

Yorio, C. A. (1989). Idiomaticity as an indicator of second language proficiency. In: K. Hyltenstam & L. K. Obler (eds.), *Bilingualism across the lifespan*: *Aspects of acquisition, maturity, and loss* (pp. 55–72). Cambridge: Cambridge University Press.

Received: 22.07.2018; **revised:** 11.09.2018