Journal of English for Academic Purposes 35 (2018) 93–104

Contents lists available at ScienceDirect

Journal of English for Academic Purposes

journal homepage: www.elsevier.com/locate/jeap

Investigating the collocations available to EAP writers

Ana Frankenberg-Garcia

University of Surrey, Guildford, GU2 7HX, UK

ARTICLE INFO

Article history:
Received 4 December 2017
Received in revised form 19 April 2018
Accepted 2 July 2018
Available online 3 July 2018

Keywords:
Collocation
Lexical competence
Academic literacy
EAP
Second language writing

ABSTRACT

Studies on the productive use of collocations have enabled researchers to harness a wealth of information about the phenomenon. However, most such studies focus on the collocations that come to the surface in finished texts, and have not been able to capture the range of collocational choices available for writers to choose from as they write. The present investigation addresses this gap by examining the collocations users of academic English at a British university were able to recall when presented with a selection of general academic writing frames. The study examined the collocations instinctively available to a group of 90 academics, tutors of English for Academic Purposes (EAP) and students at PhD, MA and undergraduate levels in an academic writing gap-filling test where more than one collocation could be used in each gap. The results indicate that experience of English academic writing plays a more decisive role than having English as a first language (L1) in the collocations effortlessly available to EAP users.

1. Introduction

It is widely acknowledged that it is not advisable for language users to learn just single words (Nattinger and DeCarrico 1992; Pawley and Syder 1983; Wray 2002). Lexical competence involves being able to put words together in texts. This requires knowledge of the world, knowledge of a language's syntactic constraints and knowledge of a language's lexical preferences. Knowledge of the world enables language users to distinguish between plausible and semantically anomalous propositions (e.g. *learn a language* versus **learn an apple*). Knowledge of a language's syntactic constraints enables users to discriminate between well-formed strings of words and errors (e.g. *give advice* versus **give an advice*). Knowledge of a language's lexical preferences, in turn, enables users to differentiate between what is acceptable, conventional and idiomatic, and what is not sanctioned by usage (e.g. *a penny for your thoughts* versus **a pound for your thoughts; *to and fro* versus **fro and to; a big mistake* versus **a large mistake*, and so on).

The latter kind of knowledge embraces a range of interconnected concepts that have confusingly come to be referred to in the literature by a variety of often overlapping terms, such as *chunks*, *collocations, fixed expressions*, *formulaic language*, *idioms*, *lexical bundles, multiword units, prefabricated units*, *set phrases*, to name but a few (Howarth 1998; Wray 2002). This is understandable, given the wide body of research from different corners of the world the phenomenon has attracted. The focus of the present study is on collocation, defined here in the Firthian (Firth, 1957) sense of 'lexical items occurring […] with a greater frequency than the law of averages would lead you to expect' (Krishnamurthy 1987, p. 70). According to this definition, collocation can be empirically verified against corpus data. It can include strings of words like *auburn hair*, where *auburn* rarely occurs in contexts other than *hair*, and *brown hair*, where *brown* occurs in many other contexts but is nevertheless still

E-mail address: a.frankenberg-garcia@surrey.ac.uk.

https://doi.org/10.1016/j.jeap.2018.07.003
1475-1585/

exceptionally frequent in the context of *hair*. Collocation can be contiguous (e.g. *carry out research*), but also allows for non-contiguous forms (e.g. *carry out much-needed research*). Collocation can cover semantically more transparent associations between words such as *cold weather*, strings that include words used in the figurative sense like *cold war*, and idioms like *cold feet*.

Appropriate use of collocations seems to facilitate comprehension. According to Hoey's (2005) Lexical Priming theory, people's minds are primed to make automatic connections between words that they have encountered together before, so word combinations that language users are already familiar with (e.g. *extenuating circumstances*) tend to be processed with less effort than combinations of words that they may not have seen before (e.g. *extenuating situation*). This view is supported by empirical studies such as Conklin and Schmitt (2007, 2012) and Ellis et al. (2008), which indicate that predicting what words are going to be used on the basis of our prior knowledge of how they normally combine facilitates language processing.

In terms of language production, the learning difficulties associated with collocations have long been acknowledged by language teachers, lexicographers and linguists. Palmer (1933, p. 5) saw collocation as 'a succession of two or more words that must be learned as an integral whole, and not pieced together from its component parts'. Hornby's *Idiomatic and Syntactic English Dictionary* – the precursor to the *Oxford Advanced Learners' Dictionary* – addressed the problem by introducing phraseological information that could help learners use words in context (Cowie, 1999). Nowadays, it is standard practice for English learners' dictionaries to provide information on collocation, and there are also specific collocation dictionaries available on the market. In the context of academic writing, resources like the *Academic Collocations List* (Ackermann & Chen, 2013) and the *Oxford Learner's Dictionary of Academic English* (Lea et al., 2014) have been compiled to cater for the particular needs of EAP learners whose first language is not English. In both cases, expert academic English corpora were used to research which target collocations to address.

Learner corpora, in turn, have been a rich source of information on learners' difficulties regarding collocations (Henriksen 2013; Paquot and Granger 2012; Wray 2013). Learner-corpus research has shown that many of the difficulties learners encounter seem to arise when collocations do not have a word-for-word equivalent in their L1. For example, Nesselhauf (2005) observed that around half the inappropriate verb-noun combinations by German learners of English could be traced back to German phraseology. Similarly, Laufer and Waldman (2011) found that the majority of English miscollocations by Hebrew learners of English were a result of literal translations from Hebrew.

However, the problem of collocations is not just one of linguistic interference leading to error. Studies such as Kaszubski (2000), Nesselhauf (2005), Durrant and Schmitt (2009), Laufer and Waldman (2011), Lu (2017), Paquot (2017) and others found that, apart from producing collocation errors, second language learners tend to prefer collocations that are congruent with collocations in their L1, and that less striking combinations of words like *very cold* are more widely used than more unique collocations like *bitterly cold*. At the same time, there is also some evidence that learners may exaggerate the use of memorable idiomatic phrases such as *far as something is concerned*, which Hasselgren (1994, p. 273) referred to as 'lexical teddy bears' because they seem like safe choices.

Despite the richness of the data generated by learner corpora and the valuable insights we have gained from them, learner-corpus research can only provide a partial picture of collocation. An important limitation is that learner corpora generally consist of a collection of short texts about a restricted set of topics that learners have been asked to write. These texts are not varied or long enough to be representative of all the collocations learners know, and the topics used to elicit the data will have influenced the collocations that surface in their writing. Of course, this problem is not exclusive to learner corpora. An elicited L1 corpus like the Louvain Corpus of Native English Essays (LOCNESS) (Granger et al. *n.d.*), which was designed to be comparable to the International Corpus of Learner English (ICLE) (Granger et al., 2002), discloses a very limited set of collocations if we compare it with the collocations present in the much larger and more varied British National Corpus (BNC), whose texts were sampled from a wide range of authentic communicative situations. Take the noun *people* as an example: despite it being the most frequent noun in LOCNESS and having a normalized frequency over four times greater than in the BNC, there is no evidence of *people* collocating with *elderly* in LOCNESS, while in the BCN there is little doubt that the two words are very strongly associated.

The above limitation is probably one of the key reasons why learner-corpus studies have revolved around the collocates of high-frequency words. For example, Kaszubski (2000) examined collocations with the verb *be*, Gilquin (2007) looked at collocations with *make*, Laufer and Waldman (2011) studied collocations surrounding the most frequent nouns in a learner corpus, and Lu (2017) examined lexical items related to the composition topics used to elicit the corpus (e.g. *pollution*). There is simply not enough data in these studies to enable one to obtain a fuller picture of the collocations learners are able to use. Moreover, as the texts that make up learner corpora are usually quite short (the texts in ICLE, for example, are between 500 and 1000 words long), many learner-corpus studies draw conclusions from what can be observed in the corpus as a whole, often overlooking individual differences between learners and idiosyncratic behaviour which could skew the data. Even if learner corpora were made up of much longer and more varied texts, however, as noted by Gilquin (2007, p. 275),

> What corpus-based studies cannot establish, […] is the extent to which collocations which are not produced by a learner are part, or not, of his/her mental lexicon. Because a learner does not produce a particular collocation does not mean that s/he does not know it (s/he may simply not need it in this specific context), but this is a side of the coin to which corpus-based approaches have no access.

To complement the information that can be gleaned from corpora, it is also possible to collect data on collocation via direct elicitation. Most of the research in this domain revolves around studies that gather introspective data to find out what

collocations language users judge as acceptable (e.g. Ellis et al., 2008; Gilquin, 2007; Granger, 1998; Wolter and Gyllstad 2011). However, the fact that language users recognize collocations does not necessarily mean they are able to use them when required. Indeed, it is generally acknowledged that language users' receptive knowledge of lexis is greater than their productive knowledge. In one of the few empirical studies that attempts to look at the two together, Laufer (1998) found that the passive vocabulary of Hebrew learners of English was much larger than their active vocabulary. Whereas Laufer's study looked at knowledge of single words, there is no reason why the same principle should not apply to learners' knowledge of collocations.

Another way of harnessing empirical data on collocation is via productive elicitation tasks, where participants are typically required to supply target words in gap-filling and/or translation tasks. For example, Gilquin (2007, p. 277) used the following gap-filling plus translation exercise to capture what verb French learners of English would combine with *choice* in the context below:

She _____ the choice of never seeing her son again.

= Elle fit le choix de ne plus jamais revoir son fils.

One of the problems of controlled tests like the above, however, is that the words participants are required to provide are not necessarily the words they would want or need to use in more naturalistic settings. Thus, unlike corpora, which 'allow learners to choose their own wording rather than being requested to produce a particular word or structure' (Granger, 2008, p. 261), gap-filling/translations tasks like the above may lack ecological validity.

Another problem is that lexical choices are not always black and white. Although gap-filling tasks are normally *designed* to elicit a single target collocation, poorly-designed tests may not always elicit the intended data (Schmitt 2010). In addition, little has been said about linguistic contexts that evoke a range of possible collocations for users to choose from. Take the noun *control* as an example. Articulate language users wishing to employ this word as an object should have little difficulty in remembering not just the word *control* in isolation, but rather collocations like *have control, take control, gain control, seize control, exercise control, exert control*, or whatever verb-noun collocations fit in with their intended meanings. In contrast, less proficient language users may have a more limited range of readily available collocations to choose from, or may simply not know what verb to use with *control*. In other words, even when language users know exactly what they want to say and what initial words to employ, a limited collocation repertoire may restrict how well they can express themselves.

However, there does not yet seem to be much research on the range of collocations available to language users at the moment of language production. As discussed above, neither learner-corpus research nor controlled gap-filling/translation tasks have so far been concerned with data that effectively taps into this aspect of lexical proficiency. In addition, there seems to be insufficient information on L1 difficulties with collocations, since existing studies tend to use L1 data as a benchmark for assessing second language performance. Yet it is important to recognize that, in the context of specific registers like academic writing, collocations could be problematic even to L1 users, who 'have to take on new roles and engage with knowledge in new ways when they enter university' (Hyland 2006, p. 2).

The present investigation is an attempt to delve more deeply into the collocation repertoire available to EAP users, where writers often struggle to put complex ideas down on paper, and where not being able to recall a suitable collocation could disrupt writing processes. More specifically, the study seeks to identify patterns in the performance of EAP users of different levels of academic experience whose first language is English (L1-English) and not English (Other-L1) in a controlled collocation test. The research questions that guided the present investigation were as follows:

1. Is the number of academic collocations available to L1-English EAP users greater than those available to Other-L1 EAP users?
2. Is the number of academic collocations available to more experienced EAP users greater than those available to less experienced EAP users?
3. Are there qualitative differences in the collocation choices by L1-English and Other-L1 EAP users?
4. Are there qualitative differences in the collocation choices by EAP users of different levels of academic experience?

2. Methodology

This section describes the participants taking part in the study, presents the elicitation materials and procedure used, and details how the data was transcribed and processed.

2.1. Participants

The participants in the study were 90 students and members of staff at the Languages Department of a British University, whose details are provided in Table 1.

The sampling was opportunistic, as the researcher worked at the Department, which facilitated the data collection. Because it was a Languages Department, it should be noted that the participants were probably more linguistically aware than average. It is also important to bear in mind that the different groups taking part in the experiment were not homogeneous in

Table 1
Participants.

Role	L1-English	Other-L1	Total
Academic	8	6	14
EAP tutor	17	3	20
PhD student	2	9	11
MA student	8	10	18
UG student	21	6	27
Total	**56**	**34**	**90**

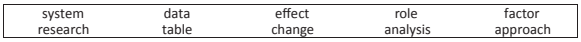

Fig. 1. Collocation bases used in the study.

terms of academic experience – defined here in terms of participant role in higher education, with the undergraduates being regarded as the least experienced group and the academics as the most experienced one[1] – or L1. The L1s other than English represented were Mandarin (8), Spanish (6), Italian (4), Polish (3), Russian (3), Greek (2), Arabic (2), Farsi (1), French (1), German (1), Slovak (1), Thai (1) and Turkish (1). The cohort was nevertheless a fair reflection of the population of the Department, and indeed of the mix of backgrounds that is often seen in British universities.

The level of English of the students from L1 backgrounds other than English met the university's entry requirements, i.e., undergraduates scored a minimum of 6, and MA and PhD students scored a minimum of 7 in the writing component of the IELTS (International English Language Testing System) or provided evidence of equivalent qualifications. The level of English of the academics and EAP tutors with L1s other than English was not formally assessed, but can be assumed to be very high, given their roles in higher education.

2.2. Materials

In order to gather data for the present study, ten nouns frequently used in general academic English served as bases for eliciting the collocations available to EAP writers. While it is recognized that there are important lexical variations in different disciplinary fields (Hyland and Tse 2007), this study takes the view supported by Coxhead (2000), Ackermann and Chen (2013), Gardner and Davies (2014) and others that there is a common core of academic vocabulary that can be useful across disciplines.

The nouns used in the experiment were selected from the Academic Vocabulary List (AVL), compiled by Gardner and Davies (2014). AVL is based on the 120-million-word academic component of the Corpus of Contemporary American English (henceforth referred to as COCA_ac). Although COCA_ac is a North-American corpus (Davies, 2008), researchers from all over the world publish in American journals, and general academic English vocabulary was felt to be sufficiently international to sanction the use of this conveniently open-access corpus as a starting point for the present study.

The ten nouns selected as collocation bases are listed in Fig. 1. They were chosen among the fifty most frequent nouns in AVL, so it can be assumed that all the participants taking part in the experiment would be familiar with them. Another important criterion in the selection of those nouns was to ensure that they could activate a range of EAP collocations rather than a single target collocation. In addition, it was determined that the nouns should evoke collocations that could be used across disciplines, rather than being specific to one particular discipline. In order to ascertain these criteria were met, the collocates of each noun were inspected in COCA_ac. For example, the noun *system* evoked adjectival collocates such as *solar, immune, new, political, legal, nervous, public, educational*, and so on. Despite the variety of adjectives retrieved, many were only attested in sources pertaining to specific disciplines. On the other hand, COCA_ac rendered a good range of verbal collocates (e.g. *implement, develop, design*) attested in different disciplinary areas within COCA_ac, which made VERB + *system* a suitable test item for the present study.

Having selected the collocation bases to be used, it was important to ensure that the elicitation task would put the participants in the right frame of mind for EAP. The nouns were thus presented within contexts pertaining to gapped academic English concordances from COCA_ac. These were piloted with two experienced academic writers, and a few adjustments were made to ensure the test items elicited the data anticipated. The sentence excerpts used are listed in Fig. 2, with examples of typical collocations from COCA_ac given in italics. Unlike traditional gap-filling collocation tests in which participants are asked to supply collocations that are not necessarily relevant to their language needs, it can be seen that the elicitation frames of the present study are typical of the kind of texts the participants encounter routinely in their work.

As shown, some of the frames in the test are more restrictive than others. For example, while in item 7 practically any adjective that collocates with *role* in academic English would be acceptable, in item 8 the missing verb needs to collocate with

[1] The study did not factor in the experience of participants who might have taken a second undergraduate, MA or PhD degree, or experience acquired in non-English academic settings.

1. The objective is to _____ a system that... *design, implement, develop*	6. Another _____ change observed was... *significant, major, dramatic*
2. Current research has _____ that... *indicated, shown, found*	7. These decisions play a/an_____ role in.... *important, key, crucial*
3. The data _____ during the process... *collected, obtained, gathered*	8. The analysis was_____ in two stages... *performed, conducted, carried out*
4. The information _____ in table 3... *provided, summarized, contained*	9. An additional factor that _____ these results was... *affected, influenced, contributed to*
5. They attempted to _____ the effect of.... *measure, examine, analyze*	10. Johns (1991) _____ a different approach to.... *adopted, used, took*

Fig. 2. Test items and example solutions (in italics).

analysis as an object and at the same time be complemented by *in two stages*. It was thus anticipated that the level of difficulty of the test items would vary. This served our purposes well, as it would help to better discriminate between more and less proficient users of academic collocations.

2.3. Procedure

The participants were told they would be presented with ten gapped sentence excerpts and were asked to fill the gaps with as many words as they remembered in the context of academic English without having to stop and think. The idea was to capture only collocations they could retrieve effortlessly, without disrupting their writing processes. It was explained that, apart from using single words, they could also supply a combination of words, such as a verb and a preposition. The participants were instructed to write a question mark (?) if they could not think of any word for a particular gap and to move on to the next sentence. This was to ensure that the gaps were not left blank because they had been inadvertently skipped, but rather because the participants were not able to retrieve a suitable word. Examples of each of these situations were given in the instructions. After reassuring the participants that the test was entirely anonymous, it was emphasized that they should not dwell on each test item, and should only supply the words that automatically came to their minds, and then move on to the next one. To ensure the test only captured lexis that was recalled effortlessly, the participants were explicitly instructed not to go back and revise their answers. These instructions were supplied in writing and explained orally. No time constraints were imposed so as not to give an advantage to faster-thinking participants over those more deliberate in their writing. What was important was to capture the moment words failed them, rather than how fast they could fill in the gaps. It took no more than 5 min for the participants to complete the task.

2.4. Data transcription

When transcribing the lexical items supplied, three illegible words could not be processed. Six spelling mistakes were corrected in the transcription, as they were not deemed relevant to an analysis focusing on lexis. However, commonly mistaken words like *affect/effect* were transcribed literally. Where participants supplied different inflections of a lemma, like *An additional factor that affects/affected these results was* …, only the first form was transcribed, since the lexical choice remains the same. Five gaps were filled in with entire phrases rather than collocations, like *An additional factor that reduced the significance of these results.* These phrases were not relevant to the study so were not transferred to the database. One EAP tutor altered the preposition preceding *two stages* in test item 8 from *in* to *into*, and filled in the gap with *divided*. This response was invalidated.

2.5. Data classification

After transcription, the words in the gaps were sorted according to whether they qualified as EAP collocations by checking them against a corpus of academic English. At this point in the study, the 37-million-word Pearson International Corpus of Academic English (PICAE) had been kindly made available to the researcher on Sketch Engine (Kilgarriff et al., 2014). PICAE is made up of texts covering a wide range of academic disciplines from American, Australian, British, Canadian and New Zealand publications (Ackermann et al., 2010), and was preferred over the admittedly larger COCA_ac because collocation look-ups on Sketch Engine are faster and more efficient, which, as shall be seen below, greatly facilitated the analysis.

It was determined that for the lexical items supplied to qualify as collocations, they had to score high in terms of strength of association and be sanctioned by a minimum number of analogous co-occurrences in different texts in PICAE. Gablasova, Brezina, and McEnery (2017) discuss different methods for establishing whether combinations of words in a corpus can be

Rank	t-score	MI	logDice
1	the	janus-faced	important
2	important	Theta	key
3	's	pivotal	central
4	their	allot	gender
5	a	caring	crucial

Fig. 3. Lemmas one-word left of *role* in PICAE ranked according to T-score, MI and logDice.

adopt +	200	9.86
propose	34	7.58
favour	18	7.32
develop	79	7.26
take +	208	7.17

Fig. 4. Word Sketch excerpt showing verbs used as objects of *approach* in PICAE, plus co-occurrence frequencies (left) and logDice scores (right).

considered collocations, including traditional strength of association measures like *t*-score and MI, and more recent ones like logDice. The association measure used in the present study was the logDice statistic favoured in Sketch Engine (Rychlý 2008). It is more robust than the *t*-score, which is overly sensitive to high-frequency words, and more appropriate than the MI-score, which rewards low frequency items, including very rare or even misspelled words. As Gablasova et al. (2017, p. 164) explain, logDice 'highlights exclusive but not necessarily rare combinations'. This is exemplified in Fig. 3, which shows the top five lemmas immediately to the left of *role* in PICAE, ranked according to *t*-score, MI and logDice.

An exploratory investigation of what could be a reasonable cut-off point in terms of logDice and co-occurrence frequencies was conducted by examining the collocates of *table* (N) and *system*, the collocation bases of the study that had respectively the lowest and highest frequencies in PICAE. In consultation with an EAP expert, a threshold of logDice $\geq$3 and a minimum of five analogous co-occurrences in at least five different sources in PICAE was found to work well for both bases. It naturally captured very frequent collocations in academic English like *develop a system*, but was at the same time sensitive to less common but nevertheless idiomatic collocations like *devise a system.* On the other hand, the cut-off point left out semantically sensible combinations of words that are arguably not appropriate in an academic register like *come up with a system*, and other plausible but more open-choice combinations like *discover a system.* As expected, it also excluded less obvious combinations of words like ?*hypothesize a system*.[2]

Sketch Engine's Word Sketch option conveniently sorts collocations according to their grammatical relations (i.e., objects, subjects, modifiers, and so on) and ranks them in terms of co-occurrence frequencies and logDice score (Fig. 4). This enabled one to validate the main collocates for each gap efficiently and flexibly, since the results are not constrained by contiguous co-occurrence or the exact wording of each test item, but allow for analogous contexts of use, as exemplified in Fig. 5.

Despite the convenience of how collocations are displayed in Sketch Engine, it was nevertheless necessary to carry out complementary concordance queries and inspect them manually in order to (1) check that the lexical items attested in Word Sketches pertained to a minimum of five different sources in PICAE, and (2) verify whether lexical items which did not figure in Word Sketches (because of parsing problems or limitations of the Sketch Grammar rules underlying them) could have nevertheless satisfied the criteria established to qualify as collocations. Note that when undertaking this analysis, spelling variants like *favour* and *favour* were considered together.

3. Results

A breakdown of the overall results is provided in Table 2. The number of blanks was very small (35), constituting only 1.5% of the total number of responses. Of the 2330 lexical items elicited in the test, 1664 (70.6%) were classified as collocations according to the criteria specified in 2.5.

[2] Note that the cut-off point of five co-occurrences (0.14 per million) is very low compared to the 1.0 per million threshold Ackermann and Chen (2013) adopted in the compilation of Academic Collocations List. However, while the aim of Ackermann and Chen (2013) was to investigate the most useful collocations to EAP learners, the objective of the present study was to establish whether a given combination of words could qualify as an academic collocation, even if not particularly frequent. The purposes of the two thresholds were thus essentially different. Moreover, raising the bar for acceptance in the present study to the one used in the Academic Collocations List would have excluded strong collocations, such as *adopt a system* (logDice = 6.65), *propose a system* (logDice = 6.69), *install a system* (logDice = 6.63), and many others. In fact, it could be argued that the threshold used in the Academic Collocations List was perhaps overly strict, since it leaves out many pedagogically relevant collocations. For example, no verbal collocates for *system* can be found in the list.

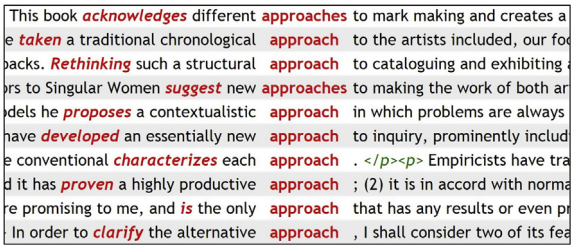

Fig. 5. Concordances feeding into Word Sketch for verbs used as objects of *approach* in PICAE.

Table 2
Overall test responses.

Responses	Lexical items	Collocations	Not Collocations	Blanks
Total	2330	1644	686	35
Median	24.50	18	7	0.00
Mean	25.9	18.27	7.62	0.4
SD	10.5	7.9	5.0	0.6
High	70	44	26	2
Low	7	2	0	0

Table 3
Test scores according to L1.

	L1-English	Other-L1
Median	19	16.5
Mean	18.89	17.24
SD	7.8	7.7
High	44	36
Low	4	2

Section 3.1 examines the 1664 elicited collocations in terms of the participants' L1 and their level of academic experience from a quantitative perspective (RQ1 and RQ2). Section 3.2 reports on the participants' collocation choices from a more qualitative perspective (RQ3 and RQ4). The 686 lexical items that did not reach the collocation threshold defined in 2.5 will be submitted to acceptability judgement testing in a follow-up study.

3.1. Quantitative findings

As previously shown in Table 2, there was considerable variability in the performance of the cohort. One participant retrieved as many as 44 collocations (averaging 4.4 per test item), while another one was only able to supply 2 collocations in the entire test. This section examines this variability from the perspectives of L1 background (RQ1) and academic experience (RQ2).

Table 3 summarizes performance according to L1 (RQ1). As shown, the L1-English group did on average slightly better in the test. Since the scores were not normally distributed, a non-parametric Mann-Whitney one-tailed test was used to determine the statistical significance of these results. With $U = 861.5$, $p < 0.05$, the difference was *not* statistically significant. This means it is not possible to rule out the possibility that the slightly higher average score of the L1-English participants was due to chance, and therefore it is not possible to make any claims about the performance of the participants on the basis of their L1.

A closer look was then taken at performance according to academic experience (RQ2). The results of this analysis are presented in Table 4. As evident in the means, there is a steady progression in the number of collocations supplied

Table 4
Test scores according to academic experience.

	Academics	PhD Students	MA Students	UG Students	EAP Tutors
Median	24.5	20	17	14	19
Mean	25.36	23.09	16.39	13.33	19.0
SD	6.8	10.5	4.3	6.5	6.0
High	41	44	23	28	29
Low	12	12	8	2	6

Table 5
Test scores by academics according to L1.

	L1-English Academics	Other-L1 Academics
Median	25	23.5
Mean	25.25	25.5
SD	8.0	5.4
High	41	36
Low	12	22

Table 6
Test scores by MA students according to L1.

	L1-English MA Students	Other-L1 MA Students
Median	18	15
Mean	17.5	15.5
SD	3.5	4.8
High	22	23
Low	12	8

that correlates with experience in academia, with the undergraduates supplying the fewest collocations, and the academics retrieving on average almost twice as many. The EAP tutors positioned themselves between the MA and PhD students, but were excluded from further comparison because their academic qualifications had not been controlled for. A one-way ANOVA was used to investigate whether the differences among the remaining participants were significant (after asserting all values satisfied the conditions of normal distribution). The results of the test were significant ($F_{33,66} = 11.79$, $p < 0.0001$), with a high effect size value ($\eta^2 = 0.65$), meaning the differences observed are substantial and unlikely to be due to chance. A post-hoc Tukey test revealed the academics significantly outperformed the undergraduates and MA students, and the PhD students significantly outperformed the undergraduates. The remaining differences (Academics vs. PhD; PhD vs. MA; and MA vs. UG) were not statistically significant. These results suggest that academic experience affects the number of collocations available to EAP users, and that the wider the gap in experience the more discernible its effect.

At this juncture, it must be remembered that the undergraduates (and EAP tutors) were predominantly L1-English speakers, while the PhD students were mostly from other L1 backgrounds. These groups were too unequal to allow for a more fine-grained analysis of the extent to which the L1 variable may have affected the results in Table 4. However, for the two remaining groups – the academics and the MA students – the distribution of L1-English and other L1s was reasonably balanced. Table 5 therefore details a comparison of the performance of the academics in terms of L1, and Table 6 summarizes analogous data for the MA students.

The values in Table 5 indicate that the differences between L1-English and other academics seemed negligible. Unsurprisingly, a one-tailed *t*-test showed the differences detected were *not* statistically significant ($t = 0.06574$, $p < 0.05$). There is therefore no evidence that having L1-English had an effect on the number of collocations the academics were able to recall.

Table 6 shows that the L1-English MA students were able to provide on average slightly more collocations than the other MA students. However, a one-tailed *t*-test indicated that the former did not significantly outperform the latter ($t = 0.9828$, $p < 0.05$). Thus, as with the results obtained for the academics, for the MA students too it was not possible to assert that having English as a first language significantly affected the number of EAP collocations remembered.

3.2. Qualitative findings

This section examines the participants' lexical preferences. As the focus was on lexis, different forms of the same lemma and spelling variations (e.g. *An additional factor that **affected**/affects these results*; *The data **analysed**/analysed during the process*) were grouped together and represented by their most frequent form (in bold in the above examples). Additionally, only lexical items favoured by at least 20% of each group were taken into account, as below this threshold there was too much idiosyncratic variation for the analysis to be meaningful.

Table 7 displays the collocations favoured by the participants in each L1 group (RQ3). Overall, there were 27 different collocations that at least 20% of the L1-English group agreed on, and 25 among the other group. Despite this slight difference in lexical diversity, it can be seen that the participants of both groups tended to use the same collocations. In eight of the ten test items, the most frequent collocation was actually the same. One noticeable difference, however, was that the L1-English participants were more prone to using high-frequency, general English lexis like *give*, *key*, *do* and *take*. In contrast, the other participants agreed more often on the use of more specialized words like *analyse* and *propose*.

When examining the lexis chosen by the participants according to academic experience (RQ4), there was more variation in the lexical preferences of each group. As shown in Table 8 the academics agreed on the greatest number of different collocations (39), which can be interpreted as an indication of a more varied and consolidated collocation repertoire. Collocation diversity correlated with academic experience, with the undergraduates agreeing on the fewest different collocations (18).

Table 7
Collocations favoured by ≥ 20% of participants according to L1 (unique items in bold).

Test item	L1-English	Other-L1
1	Create Develop	Create Develop
2	Shown Proven **Suggested** Demonstrated Found	Shown Proven Demonstrated Found
3	Collected Gathered	Collected **Analysed** Gathered
4	Shown **Given** Provided Presented	Shown Provided Presented
5	–	**Analyse**
6	Significant Important	Significant Important
7	Important Significant Vital Crucial **Key**	Important Significant Vital Crucial
8	Carried out **Done** Conducted	Carried out Conducted
9	Affected Influenced	Influenced Affected
10	Suggests **Takes**	Suggests **Proposes**
Total	27	25

Interestingly, the EAP tutors positioned themselves between the academics and the PhD students on this scale. The most frequent collocations chosen by each group coincided in only five out of ten test items (items 2, 3, 4, 7 and 10), suggesting there is again more variability in relation to academic experience than in terms of L1 background. Another interesting finding is that the academics agreed on the greatest number of collocations that were unique to their group (13), which could be another indication of a more consolidated collocation repertoire. In contrast, none of the collocations agreed upon by at least 20% of the undergraduates were unique to their group.

4. Discussion and conclusion

It has long been acknowledged that lexical knowledge is not just about understanding words, but also about employing words in context. Corpora have enabled researchers to capture how linguistic communities conventionally put words together, and learner corpora have brought to light problems that are typical among less proficient language users. However, corpora cannot provide information on the lexical choices available to writers at the moment of writing. The present study set out to investigate the collocations available to a group of 90 EAP users in a controlled language production task designed to elicit academic collocations.

The lexical items the participants supplied varied in number and in type. This is not unexpected, as the study cohort was not and should not be treated as a homogeneous group. Although the participants were all from a Languages Department, and therefore likely to be more linguistically aware than the average EAP user, their uneven performances serve to underscore the fact that there can be substantial variability in the productive collocational repertoire of regular users of academic English. This heterogeneity should be acknowledged and understood.

One factor that could explain the differences observed is that many EAP users do not have English as a first language. However, no significant differences were found in the number of collocations available to participants with and without L1-English. Even though the initial overall results could have been distorted by the fact that the L1 groups in the opportunistic experimental cohort were not balanced, when the two subgroups that were similar in size were compared separately, the results were the same. These quantitative findings were reinforced by the qualitative analysis that followed. The participants of both language groups tended to favour the same collocations. One small but noticeable difference detected, however, was

Table 8
Collocations favoured by ≥ 20% of participants according to academic experience (unique items excluding EAP-tutor performance in bold).

Test item	Academics	PhD Students	MA Students	UG Students	EAP Tutors
1	Develop **Design** Create	Develop Create **Devise** **Describe** Set up	Create Provide Develop Establish	Create	Create Design Develop Set up
2	Shown Proven Demonstrated Suggested **Indicated** **Established**	Shown Proven Demonstrated	Shown Proven Found Demonstrated Suggested	Shown Proven Suggested	Shown Proven Found Indicated Demonstrated
3	Collected Gathered Analysed	Collected Analysed	Collected Gathered Analysed **Found**	Collected	Collected Gathered Found
4	Shown Displayed Presented Provided **Contained**	Shown Provided Displayed	Shown Provided Given	Shown Given Presented Provided	Shown Presented Given Displayed Provided
5	**Show** **Demonstrate** Analyse	**Explain**	Analyse	–	Analyse
6	Significant **Important** **Interesting**	Significant **Major** **Unexpected**	–	Significant	Significant Important Noticeable Interesting
7	Important Significant Crucial Key Vital	Important Significant Key Crucial **Major** **Minor** Vital	Important Vital Significant	Important Key Essential Crucial	Important Significant Vital Crucial Essential Major
8	Carried out Conducted **Developed** Done	Conducted Done	Carried out	Done Carried out	Conducted Carried out Done
9	Affected **Impacted on** **Contributed to**	Affected Influenced **Led to**	Affected Influenced	Affected Influenced	Influenced Affected
10	Suggests **Takes** **Developed** Proposes	Suggests	Suggests Proposes	–	Suggests Takes
Total	39	30	25	18	35

that the collocation repertoire of the L1-English participants tended to be more permeable to less formal, general English lexis. This could be explained by the fact that they will normally have had more exposure to non-academic uses of English.

The main variable affecting the number and variety of collocations available to the participants in the study was their level of academic experience. The undergraduates supplied the fewest collocations, the MA students came next, then the PhD students, and finally the academics. Additionally, the EAP tutors outperformed the MA students and undergraduates.

Of course, the positive correlation of collocation repertoire and years at university could have been skewed by the imbalanced L1 backgrounds pertaining to different levels of academic experience in the cohort. However, as pointed out above, when the two reasonably balanced groups were compared, their performances did not differ significantly. Moreover, when considering the cohort as a whole, it should be noted that the Other-L1 bias was stronger among the PhD students (with more academic experience), while the L1-English bias was more pronounced among the undergraduates (with less academic experience). Therefore, it cannot be inferred that the groups of higher academic experience performed better because there were more L1-English participants among them, and neither that the groups of lower academic experience did less well because there were more participants whose first language was not English among them. If anything, quite the opposite was true.

These findings indicate that having English as a first language does not automatically give an advantage to users of academic English in terms of their productive collocation repertoire, and lend support to the view that there are no native speakers of Academic English (Hyland 2006; Hyland and Shaw 2016; Kosem 2010). Moreover, in line with Hulstijn's (2011) theory of Higher Language Cognition, the present findings suggest that L1 performance should not be indiscriminately used as a benchmark for assessing L2 proficiency, particularly when dealing with a specialized register like EAP.

The qualitative analysis showed the academics not only supplied more collocations, but were also more consistent in their lexical choices. They supplied more collocations in common than the other groups, particularly the undergraduates. These findings are consistent with Hoey's (2005) Lexical Priming theory, whereby language users take mental notes of how words are used, and learn to make automatic connections between such words once they have encountered them together sufficiently often. In the present case, years of experience in reading and writing academic texts seem to have equipped the academics with a more sophisticated and more consolidated collocation repertoire, regardless of their having L1-English or not.

In terms of implications for teaching, the present findings suggest that novice EAP users would benefit from further awareness of and exposure to academic collocations, even when their L1 is English. While extensive reading and writing over the years at university appears to be an effective way of boosting one's collocation repertoire incidentally, it is important to recognize that the writing of novice EAP users may be disrupted by a less than optimal recall of academic collocations, and that L1-English writers too may need support in this respect. Dictionaries and other collocation resources can jog writers' memories at the moment they need a specific collocation, and it would not be unreasonable to suggest that EAP collocation references like the Academic Collocations List appended to the *Longman Collocations Dictionary* and the *Oxford Learner's Dictionary of Academic English*, which have been compiled specifically for EAP users of L1s other than English (Ackermann & Chen, 2013; Lea 2014) could in fact also be useful to L1-English undergraduates and secondary school students who have not yet had sufficient opportunities to assimilate the lexical conventions of the register.

This does not mean to say that differences in L1 background should go unacknowledged. There is abundant evidence on the negative impact of incongruent L1/L2 collocations (e.g. Laufer and Waldman 2011; Nesselhauf 2005; Peters 2016). Moreover, the present study generated new data indicating that L1-English EAP writers tend to be more prone to using general English lexis in academic contexts. The planned follow-up investigation where the 686 non-collocations elicited in this study will be subjected to acceptability testing should disclose further insights about the effect of L1. Questions such as whether the words classified as non-collocations were open-choice, errors, too informal or just odd in an EAP context remain to be answered.

Another issue that should not be overlooked is the difference between core and discipline-specific collocations. Although the present study did not examine discipline-specific collocations, it pointed to deficiencies in the use of core collocations by novice EAP users. This lends credibility to the pedagogical value of EAP vocabulary resources that cut across different academic domains proposed by Coxhead (2000), Gardner and Davies (2014), Ackermann and Chen (2013), Lea (2014) and others. In fact, one must not discard the possibility that general EAP collocations might well be harder to acquire incidentally, since they could be less noticeable when compared with the more targeted and concentrated way in which EAP users are exposed to discipline-specific collocations.

Having said this, the present findings are exploratory and should be interpreted with caution. A larger investigation, with a more balanced cohort in terms of L1 and academic experience, and that includes participants from different disciplinary areas, is still needed. In future, a computer-delivered test with screen recording would also enable one to better control for the exact moment collocations stop flowing. Notwithstanding these limitations, the study offers important insights into the collocations effortlessly available to EAP writers, and opens the way for further studies. Future research could usefully explore how well writers of different L1s recall congruent and incongruent collocations, and whether there are differences in discipline-specific and core academic collocation recall.

Acknowledgements

I would like to thank the three anonymous reviewers of this paper for their insightful and constructive feedback. Part of this study was conducted in the scope of the ColloCaid project, funded by the UK Arts and Humanities Research Council (AHRC) (AH/P003508/1).

References

Ackermann, K., & Chen, Y. (2013). Developing the academic collocations list (ACL) – a corpus-driven and expert-judged approach. *Journal of English for Academic Purposes, 12*, 235–247.

Ackermann, K., De Jong, J., Kilgarriff, A., & Tugwell, D. (2010). *Research summary*. The Pearson International Corpus of Academic English (PICAE). http://pearsonpte.com/wp-content/uploads/2014/07/RS_PICAE_2010.pdf. (Accessed 23 March 2017).

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

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics, 32*, 45–61.

Cowie, A. (1999). *English dictionaries for foreign learners: A history*. Oxford: Oxford University Press.

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

Davies, M. (2008). The corpus of contemporary american English (COCA): 520 million words, 1990-present. http://corpus.byu.edu/coca/. (Accessed 22 February 2017).

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL - International Review of Applied Linguistics in Language Teaching, 47/2*, 157–177.
Ellis, N., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *Tesol Quarterly, 42/3*, 375–396.
Firth, J. (1957). A synopsis of linguistic theory, 1930–1955. In J. Firth (Ed.), *Studies in linguistic analysis. Oxford* (pp. 1–32). Blackwell.
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.
Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35/3*, 305–327.
Gilquin, G. (2007). To err is not all: What corpus and elicitation can reveal about the use of collocations by learners. *Zeitschrift für Anglistik und Amerikanistik, 55/3*, 273–291.
Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 145–160). Oxford: Clarendon Press.
Granger, S. (2008). Learner corpora. In A. Lüdeling, & M. Kytö (Eds.), *Corpus Linguistics: An international handbook* (pp. 259–274). Berlin: Walter de Gruyter.
Granger, S., Dagneaux, E., & Meunier, F. (Eds.). (2002). *International corpus of learner English*. Louvain: Presses Universitaires de Louvain. http://www.uclouvain.be/en-cecl-icle.html. (Accessed 8 August 2016).
Granger, S., C. Sanders and U. Connor. n.d. LOCNESS: Louvain Corpus of Native English Essays. https://www.uclouvain.be/en-cecl-locness.html [08/08/2016].
Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics, 4*, 237–258.
Henriksen, B. (2013). Research on L2 learners' collocational competence and development: A progress report. In C. Bardel, B. Laufer, & C. Lindqvist (Eds.), *L2 vocabulary acquisition, knowledge and use. New perspectives on assessment and corpus analysis*. Eurosla Monographs Series, 2. EUROSLA.
Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London and New York: Routledge.
Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics, 19/1*, 4–44.
Hulstijn, J. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly, 8/3*, 229–49.
Hyland, K. (2006). *English for academic purposes*. London and New York: Routledge: An Advanced Resource Book.
Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *Tesol Quarterly, 41/2*, 235–253.
Hyland, K., & Shaw, P. (2016). Introduction. In K. Hyland, & P. Shaw (Eds.), *The routledge handbook of English for academic purposes* (pp. 1–14). London: Routledge.
Kaszubski, P. (2000). *Selected aspects of the lexicon, phraseology and style in the writing of polish advanced learners of English: A contrastive, corpus-based approach*. PhD thesis. Poznán: Adam Mickiewicz University http://www.staff.amu.edu.pl/~przemka/rsearch.html#PhD. (Accessed 8 August 2016).
Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovvář, V., & Michelfeit, J. (2014). The Sketch Engine: Ten years on. *Lexicography, 1*, 7–36.
Kosem, I. (2010). *Designing a model for a corpus-driven dictionary of Academic English*. PhD Thesis. Birmingham, UK: Aston University http://publications.aston.ac.uk/14664/1/Kosem2010_484017_3.pdf. (Accessed 3 December 2017).
Krishnamurthy, R. (1987). The process of compilation. In J. Sinclair (Ed.), *Looking up. An account of the cobuild project in lexical computing*. London and Glasgow: Collins ELT.
Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics, 19/2*, 255–271.
Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61/2*, 647–672.
Lea, D. (2014). Making a Learner's dictionary of academic English. In A. Abel, C. Vettori, & N. Ralli (Eds.), *Proceedings of the XVI EURALEX*. Bolzano/Bozen: Institute for Specialised Communication and Multilingualism.
Lea, D., Bull, V., Webb, S., & Duncan, R. (2014). *Oxford Learner's dictionary of academic English*. Oxford: Oxford University Press.
Lu, Y. (2017). *A corpus study of collocation in Chinese learner English*. London: Routledge.
Nattinger, J., & DeCarrico, J. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.
Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam and Philadelphia: John Benjamins.
Palmer, H. E. (1933). *'Second interim report on English collocations', tenth annual conference of English teachers under the auspices of the institute for research in English teaching*. Tokyo: Institute for Research in English Teaching.
Paquot, M. (2017). L1 frequency in foreign language Acquisition: Recurrent word combinations in French and Spanish EFL learner writing. *Second Language Research, 33*, 13–32.
Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics, 32*, 130–149.
Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory: Native-like selection and native-like fluency. In J. Richards, & R. Schmidt (Eds.), *Language and communication* (pp. 191–226). New York: Longman.
Peters, E. (2016). The learning burden of collocations: The role of interlexical and intralexical factors. *Language Learning, 20/1*, 113–138.
Rychlý, P. (2008). A lexicographer-friendly association score. In *Proceedings of recent advances in slavonic natural language processing* (pp. 6–9). RASLAN.
Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Basingstoke: Palgrave Macmillan.
Wolter, B., & Gyllstad, H. (2011). Collocational links in the L2 mental Lexicon and the influence of L1 intralexical Knowledge. *Applied Linguistics, 32/4*, 430–449.
Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.
Wray, A. (2013). Formulaic language. *Language Teaching, 46*, 316–334.

ANA FRANKENBERG-GARCIA is Reader in Translation Studies at the University of Surrey. Her research focuses on applied uses of corpora in writing, lexicography and translation. She is principal investigator of the ColloCaid project. Ana was also principal investigator of the open-access COMPARA parallel corpus of English and Portuguese fiction and chief editor of the Oxford Portuguese Dictionary.