Formulaic sequences involving 'fact' in EAP production: A corpus study[1]

Magdalena Zinkgraf

María Angélica Verdú

Facultad de Lenguas, Universidad Nacional del Comahue

maguizinkgraf@gmail.com

1. Formulaic sequences in EAP written production

A 'formulaic sequence'(Wray, 2002) was originally defined as "a sequence, continuous or discontinuous, words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (p. 9). Irrespective of the number of different definitions that have since been put forward to redefine the term, Ellis (2008) and Wood (2010) coincide that they are central to the expression of concepts and textual relationships in efficient and effective communication. Evidence of the pivotal role formulaic sequences play is the vast repertoire of memorized, prefabricated strings present in native speaker language (Erman & Warren, 2000).

Recently they have gained more importance in EFL since their acquisition by non-native language learners has been shown to present difficulties which may be a hindrance to the development of fluency in natural language use (Coxhead, 2008; Schmitt, 2010; Schmitt, Dörnyei, Adolphs & Durow, 2004).

Previous research into the importance of these strings of words in EAP has shed light on different sequences used by learners or 'apprentice writers' (Römer & Arbor, 2009) as

compared to both native and non-native expert writers' use. Studies like those by Biber (2006), Henderson and Barr (2010), Neff van Aerstselaer (2008), and Oshima and Hogue (2004) claim that EAP learners' writings display certain characteristics that are not altogether present in expert academic journals.

Many researchers have used corpus-based linguistics to compare native speakers' versus non-native learners' use of strings to identify differences between these two populations (Nesselhauf, 2003, 2005; Siyanova & Schmitt, 2008). The present paper is one such case.

2. The study

The purpose of this study is to characterize how the noun "fact", typical of academic writing contexts (Barry, 2011; McCarthy & O'Dell, 2008), is employed in formulaic sequences (FSs) involving clusters like *the fact that, a well-known fact* and *the fact is that*, and compare these expressions to instances found in the *Corpus of Contemporary American English* or *COCA* hereafter (Davies, 2008). This study will analyse the uses of these FSs in *COCA*'s academic subcorpus, which will constitute the 'expert' corpus of reference (EC).

To this end, a database of 237 written texts (114,514 words in 5,176 sentences) was studied by means of *Wordsmith Tools 6*.0 (Scott, 2015) in search of the most frequent FSs involving *fact*. Learners from cohorts between the years 2008 – 2014 were asked to write 300-to-500-word essays (mean 358.54, Std: 202.28) about any topic of a list of thesis statements provided. The first practical assignment was used to compile the learner corpus (LC).

Participants in the study were 237 EFL students who took the annual subject English IV, for the Teacher-Training and Translation Courses at Facultad de Lenguas (Universidad Nacional del Comahue) and who gave a written consent for their written production to be used for research purposes. Their age varied between 21 and 34. The LC represents a cross-sectional sample of the initial written performance of learners taking this course in this particular EFL university setting.

Following Gillett (2011)'s definition of EAP in terms of Robinson's (1991) criteria for ESP, their production is considered as belonging within the realms of EAP because of a) the university setting in which they acquire the target language, b) the type of courses they are taking (teaching and translation), and c) the text-types and task-types they are required to submit during the course.

2.1. Analysis of formulaic sequences with 'fact'

Wordsmith Tools 6.0's (Scott, 2015) was predetermined to identify recurrent clusters made up of between three to eight words in the learner corpus. Due to the size of the corpus, the frequency required was established as three or more times (Scott, 2001). Those involving *fact* were manually selected. Concordance lines were sought using the software's tool *concordance* and frequent FSs around *fact* were identified, and their number of instances, summarized in Table 1 below.

Clusters in the learner corpus	N
the fact that	115
it is a well-known fact	10
the fact of	10
the fact is that	5
Total	140

Table 1. Number of instances per cluster involving fact in the learner corpus.

In section 3, we analyse the 140 concordance lines obtained and present longer formulaic sequences recurrently used by these learners.

3. Results and discussion

The detailed study of the concordance lines in LC obtained for each of the clusters has led to a characterization of their use. In Table 2 the occurrences of each of those clusters (N) are presented for both the LC and the EC. In the latter, we have included those appearing in the academic subcorpus and in the general *COCA* (Davies, 2008), given that a more accurate comparison may highlight certain patterns of use and reveal similarities and differences between learner and expert use[2].

3.1. *the fact that*

This 3-word string, which is by far the most frequent in LC, participates in much larger sequences which appear quite recurrently. As illustrated in Table 2, the first eight FSs are the most recurrent in this corpus –shaded in white in the table- and, for all of them, learner use seems to imitate expert use as evinced in the high number of their occurrences in *COCA* Academic. Examples from the learner corpus are offered below for the causative function (1) and (2), *due to the fact that* and *owing to the fact that*. The use of the latter, though much less frequent in the general *COCA* and in the academic subcorpus, might indicate a slight degree of overuse on the part of these learners.

(1) at in many cases they keep up the pretence **due to the fact that** they have fear of speaking publicly

(2) use them without disturbing other people **owing to the fact that** we can use headphones which are usef

Formulaic sequence	Learner corpus (114,514 words)	*COCA* (450 million words)	
	N	N	
		General	Academic (within general)
the fact that	**115**	**49860**	**12822**
due to the fact that	31	633	372
owing to the fact that	3	28	8
despite the fact that	6	2627	98
in spite of the fact that	5	362	127
(be) aware of the fact that	7	262	57
conscious of the fact that	1	65	20
Has/have to do with the fact that	4	113	50
related to the fact that	3	83	51
connected to the fact that	1	8	3
given the fact that	1	689	172
stems from the fact that	0	200	100
results from the fact that	1	21	15
illustrated by the fact that	1	27	22
shown by the fact that	1	19	14
consider the fact that	2	111	27
considering	1	100	23
bearing in mind the fact that	1	2	1
bear in mind	1	4	0
it is a well-known fact that	**11**	**46**	**4**
the fact of	**10**	**2609**	**370**
the fact is that	**3**	**1977**	**220**

Table 2. Comparison between instances of FSs in the learner corpus and COCA.

Two other FSs which present a contrast are used by learners: both *in spite of the fact that* (3) and *despite the fact that* (4) appear a similar number of times in LC (6 and 5

occurrences, respectively), resembling expert use in the academic subcorpus of *COCA* (cfr. 127 and 98 each).

(3) of our country, it has been going on **in spite of the fact that** mining laws forbid it. As it seems t

(4) are what make us good and decent people. **Despite the fact that**, nowadays, everyone seems to look at

The formulaic sequence *(be) aware of the fact that* is exemplified in (5), together with its variation *(be) conscious of the fact that*, both of them used in similar proportions as those found in *COCA*, a sign of learner sensitivity to expert preference.

(5) on the contrary, a lot of consumers **are aware of the fact that** most of the products they buy are ma

In (6) the FS *has/have to do with the fact that* is exemplified to offer a connection between two factors or behaviours and as an explanation for these phenomena.

(6) One of the main issues of this problem **has to do with the fact that** people, in this case women, are constantly exposured (*sic*) to daily TV programmes, fashion m

Learner use of this FS mirrors that of native speakers. Similar in function are two other sequences encountered in LC: *(be) related to the fact that* (7) and *(be) connected to the fact that*. The latter is neither frequent in LC or *COCA*, and in fact, learners should perhaps be discouraged from using it altogether.

(7) Another reason why it is important to have a counselor at university **is related to the fact that** students become more independent when starting university

Three infrequent FSs in LC present a resultative expression whose use should be encouraged due to the high number of occurrences in *COCA* academic (shaded in light grey in Table 2): *given the fact that*¸ *result from the fact that,* and *stem from the fact that*. While there is one instance of each of the first two in LC, illustrated in (8) and (9) respectively, the third is entirely absent from it, while particularly frequent in the academic subcorpus of *COCA*.

(8) le for creating a bad influenced on society **given the fact that** most people have damaged their own h

(9) rcise nor eat healthy. These problems **result from the fact that** the vast majority of students avoid

These FSs could perhaps be offered as alternatives to the very frequent *due to the fact that*, and thus make learners' written production richer, more varied and adequate for EAP writing standards.

In order to present examples of the issues raised in their essays, learners have resorted to an instance of each of the following FSs, imitating expert use as evinced in *COCA*:

(10) ce on adolescents. **This is clearly illustrated by the fact t**hat young girls are at an age in which t

(11) super thin means being healthy. **This is shown by the fact that** the members of the fashion industry

Another typical FS, which is quite recurrent in *COCA* and which appears in LC, is the frame that involves *[adj](be) the fact that*. Learners have made attempts at using these strings as shown in (12) and (13) below:

(12) y difficulties. **Equally relevant to this issue is the fact that** children are also negatively influen

(13) object to be perfected. Perhaps **most alarming is the fact that** media images of female beauty are un

These FSs are indicators of a more developed formulaic competence because of the modification in the canonical word order they display, In *COCA*, however, this frame is introduced by adjectives like *significant, surprising, disturbing, disconcerting,* and *remarkable,* among others, whose acquisition should be fostered in learner academic writing.

Some other FSs that introduce the 3-word cluster present in LC and frequent in *COCA* are *highlight the fact that; hold true to the fact that* and *take issue with the fact that*, of which one example has been encountered. In the expert subcorpus these FSs appear with a very high frequency rate (65, 69 and 115 respectively), which might point to the need to include them in learners' repertoire of academic formulaic sequences.

Certain differences discovered between LC and *COCA* academic are related to what learners have actually *not* used in the company of *the fact that*. The most frequent collocates, apart from the typical function words, in the expert corpus for this cluster include verbs such as *reflect, ignore, evidenced, derive* and *reinforce* in their varied forms, and none of them occurs in LC. Instruction on the use of such verbs would benefit EAP learners.

Among the few divergences found between learner and expert FSs involving the three-word cluster under study, we have discovered the inadequate use of *bear in mind the*

fact that, which is scarce not only in the academic subcorpus but also in the general *COCA* (see Table 2).

The comparison between corpora has also shed light on some problematic issues such as mistakes in the LC related to subject position of the cluster under analysis, generated by anticipatory 'it', as in of (14).

(14)*First of all, **it must be considered the fact that** (*sic*) learners are not isolated. They are parts of groups of students that attend to the sam

Even if this type of error is typical of EFL learners (Hewings and Hewings, 2002), it is precisely the insertion of the cluster in question that is conducive to this type of mistake.

3.2. *It is a well-known fact that*

This FS appears overwhelmingly frequently in LC (11 occurrences). In comparison with the academic subcorpus of *COCA*, this is the only case of the FSs analysed where there are more instances in the learner than in the expert corpus. This phenomenon probably shows overuse on the part of learners, who most likely have perceived this FS and its frequency in the input while unaware that, as part of its restrictions, it is not typical of academic discourse, and have thus generalized its use to cases where it is not common.

3.3. *the fact of*

This is a quite frequent FS in LC and its use reflects that in *COCA*. Learners, as well as native speakers resort to it to introduce the subject of their utterance as illustrated in (15)

(15) ls on TV create on normal-sized women. Sometimes, **the fact of** wanting to be very slim becomes a dang

3.4. *The fact is that*

Contrary to findings in *COCA* academic, learner use of this FS is quite infrequent. Though three instances have been found in LC, it might be advisable to draw EAP learners' awareness to the recurrence of this sequence.

4. Conclusions

This descriptive study has investigated some FSs typical of academic discourse which learners at this stage are actually familiar with and use productively. It has also presented evidence of overuse of FSs they take to be frequent, failing to reflect expert use. The analysis has further unveiled mistakes triggered by the use of some of these clusters. The information obtained can serve as a compass to guide teachers in this particular context in their selection and teaching of typical FSs in EAP.

One word of caution is needed here as regards the conclusions that can be drawn from this study. The findings derived from the comparison between COCA and our learner corpus may provide us with a few guidelines as to how to develop learners' formulaic competence further. These two databases are not entirely comparable because of the differences in their make-up and size. They do, however, point to similarities and divergences from expert use, which should inform the instruction of FSs in EAP. Authors seem to agree that learners' attention needs to be drawn towards larger strings of words surrounding lexical items, to how recurrent they are in expert texts (Römer and Arbor, 2009) and to which restrictions operate in their possible uses. Through this awareness-raising process, learners can be adequately equipped to produce them in their typical linguistic environments, approximating expert language use (Flowerdew, 2001; Granger and Meunier, 2008; Wood, 2002).

Since this is a cross-sectional, exploratory study that captures four cohorts of language learners' written production at one static point in time and given that the essays compiled were the first in the academic year, further research should explore these students' development of their formulaic competence across time.

Notes

1 This paper is based on the partial analysis of database resulting from the study carried out by Julieta Pérez, the student member of research project J023, "Secuencias formulaicas y su adquisición en estudiantes universitarios de Inglés como Lengua Extranjera", subsidized by Secretaría de Ciencia y Técnica, Universidad Nacional del Comahue.

2 Cells in white indicate coincidence in frequency; light grey highlights FSs which should reinforced through instruction; dark grey signals divergence between corpora.

References

Barry, M. (2011). *Steps to academic writing*. Cambridge: CUP.

Biber, D. (2006). *A corpus-based study of spoken and written registers.* Amsterdam: John Benjamins Press.

Coxhead, A. (2008). Phraseology and English for academic purposes: Challenges and opportunities. In Meunier, F. and Granger, S. (Eds.) *Phraseology in foreign language learning and teaching*. (pp. 149-161) Amsterdam: John Benjamins Press.

Davies, M. (2008).*The Corpus of Contemporary American English: 450 million words, 1990-present*. Retrieved from http://corpus.byu.edu/COCA/

Ellis, N.C. (2008). Phraseology: The periphery and the heart of language. In Meunier, F. & Granger, S. (Eds.). *Phraseology in foreign language learning and teaching*. (pp. 1-13) Amsterdam: Philadelphia: John Benjamins Press.

Erman, B and Warren, B. (2000). The idiom principle and the open-choice principle. *Text, 20*, 29-62.

Flowerdew, L. (2001). The exploitation of small learner corpora in EAP materials design. In Ghadessy, M., Henry, A. and-Roseberry, R.L. (Eds.). *Small corpus studies and ELT: Theory and practice*. Amsterdam: John Benjamins Press.

Gillett, A. J. (2011). What is EAP? Retrieved from http://www.uefap.com/bgnd/

Granger, S. and Meunier, F. (2008). Phraseology in language learning and teaching: Where

to from here? In Meunier, F. and Granger, S. (Eds.) *Phraseology in language learning and teaching* (pp. 247-252). Amsterdam: John Benjamins Publishing Company

Henderson, A. and Barr, R. (2010). Comparing indicators of authorial stance in psychology students' writing and published research articles. *Journal of Writing Research, 2*(2), 245-264.

Hewings, M. and Hewings. A. (2002). 'It is interesting to note that …': A comparative study of anticipatory 'it' in student and published writing. *English for Specific Purposes, 21,* 367-83.

McCarthy, M. and O'Dell, F. (2008). *Academic vocabulary in use*. Cambridge: Cambridge University Press.

Neff van Aertselaer, J. (2008). Contrasting English-Spanish interpersonal discourse phrases in phraseology. In Meunier, F. and Granger, S. (Eds.). *Phraseology in foreign language learning and teaching* (pp. 85-100). Amsterdam: PA, John Benjamins Publishing Company.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics, 24*(2), 223-242

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins Publishing Co.

Oshima, A. and Hogue, A. (2004). *Writing academic English* (4th ed.). Montreal, QC: Pearson.

Robinson, P. (1991). *ESP today: A practitioner's guide*. London: Prentice Hall.

Römer, U. and Arbor, A. (2009). English in academia: Does nativeness matter? *Anglistik. International Journal of English Studies, 20*(2), 89-100.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. London: Palgrave Macmillan.

Schmitt, N., Dörnyei, Z. , Adolphs, S. and Durow, V. (2004). Knowledge and acquisition of formulaic sequences. In Schmitt (Ed.). *Formulaic sequences: Acquisition, processing, and use* (pp. 55-86). Amsterdam: John Benjamins Press.

Scott, M. (2015). *WordSmith tools version 6.0.* Liverpool: Lexical Analysis Software

Scott, M. and Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins Press

Siyanova, A. and Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *The Canadian Modern Language Review/La Revue canadienne des langues vivantes, 64*(3) (March/mars), 429–458.

Wood, D. (2002). Formulaic language in acquisition and production: implications for teaching. *TESL Canada Journal*, *20*(1), 1-15.

Wood, D. (2010). Lexical clusters in an EAP textbook corpus. In Wood, D. (Ed.) *Perspectives in formulaic language: Acquisition and communication* (pp. 88-106). New York: Continuum Books.

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge: Cambridge University Press.