10

FROM CORPUS TO CALL

The use of technology in teaching and learning formulaic language

Tom Cobb

Introduction

Linguists discovered formulaic language (FL) through computer analysis of large texts, and this chapter makes the case that second language (L2) learners should follow in their footsteps, though probably with more learner-oriented or CALL (computer-assisted language learning) types of software. Non-computational approaches to FL do not deal adequately with what is known about FL (the extent, distribution, or true nature of it) nor its acquisition (that it requires both awareness and massive exposure). While a CALL approach in this area is yet to be extensively developed or conclusively tested, this chapter will furnish concrete ideas for why and how this can be achieved and will report on progress and prospects.

Linguists and learners

It is reasonable there should be a role for computer technology in the teaching and learning of FL, because linguists themselves learned of the existence and extent of FL by looking at large texts, or corpora, with computer technologies. Prior to corpus analysis, it could be assumed that, apart from a relatively small number of idioms and set expressions ("How are you?" or "Good morning"), languages consisted mainly of words and rules – individual words assembled into phrases and sentences through the application of grammatical rules. What corpus analysis revealed, however, was that of the infinite number of phrases that can be assembled by the application of rules to words, only a few of these are actually used by speakers of any language. Furthermore, this usage is shared by all users of a language and is an important part of what we know when we know a

language. In other words, there is another principle in addition to grammar that is operating in the construction of any utterance. It is what Sinclair (1991) called "the idiom principle", though it applies equally to literal and metaphorical or idiomatic expressions. Of all the many grammatically permissible ways of proposing marriage to someone ("Would you consider marriage? Does your marrying me seem plausible?" etc.), the precise formula "Will you marry me?" claims the vast majority of instances – whether in linguistics, literature, or life – and learners should know this.

The idiom principle was suspected to exist by many in the pre-computational era of linguistics (Pawley & Syder, 1983), but what was new with corpus analysis was its extent. Starting in the mid-1980s, computer software was written that could analyze texts of several million words, tallying among other things the amount of word-group recurrence, and this turned out to be unexpectedly large. A classic finding from Erman and Warren (2000) is that 52% of word tokens in typical spoken text and 40% in written text are involved in some sort of word-group recurrence. A further twist is that if one or two intervening words are allowed to count as part of the group ("big *shiny* car" as well as "big car") and members of word families are counted as repeated words ("big shiny cars") then these figures can rise by another 15%. (Explore this claim at N-Gram on the author's *Lextutor* Website, http://lextutor.ca/n_gram/, by manipulating any text with the "Intervenors" and "Families" settings.)

Admittedly, there is a question about the true formulaicity of computer-generated phrase repetitions or "bundles". Is the recurrent string "of the many" a meaningful unit? Such items must form a large part of Erman and Warren's figure. This question has led to a reaction against pure computational approaches by the phraseologists (e.g., Cowie, 1998), who propose either intuition, hand work, or hand checking of computer work as the most reliable source of insight about FL. The fact remains, however, that the extent of formulaicity in language was first learned of through computational analysis, and this remains an awareness-raising insight that is important for language learners to experience, so in principle such analysis can also be a source of insight for them, too. But does the argument for computation in learning about formulaic language extend beyond "in principle"?

One reason for believing so is that while computer analysis of a text or corpus may occasionally focus a learner's attention on a non-formulaic string like "of the many", this at least is a true recurring word group which will cause no harm if attended to, as cannot be said for those proposed in some other approaches to teaching and learning formula. Boers, Dang, and Strong (2017) found that most formula exercises across 10 current EFL textbooks invited learners to fill in blanks with suitable items from memory or imagination, or else make matches on a table between, e.g., *drive/ride* and *bike/car*, either of which left many with a strong memory for precisely the non-standard association, possibly because of the processing effort committed to raking memory and/or guessing. Non-standard associations simply could not be produced from the consideration of concordance

lines, or any pedagogical exercise derived from these, for the simple reason that they are not present in that output. Figure 10.1 shows a selected concordance from a graded reader corpus for the keyword *ride* with typical collocates italicized by a teacher (and hence a hand-checked computer output, in terms of the prior discussion). The concordance lines bear nothing resembling "ride a car" but rather bear numerous words that typically do harmonize with *ride*, like *horse*, *bicycle*, and *bus*. In other words, the pedagogical case for using corpus technology to learn formulaic sequences (FSs) at least extends as far as providing little or no unsuitable learning stimuli, or "doing no harm" (in the words of a well-known formula).

The ways a concordance interface can explore the FL of a text or corpus are basically two. First is a search with hypothesis in hand, which consists of entering an intact phrase ("bus ride") with the options of intervening items, either item first, or alternative morphologies. Second is an exploratory search, which consists of entering a keyword ("ride") in one or all morphologies, with the option to sort adjacent words by first, second, or third word to the right, or left – one of these sortings should expose any repeated associations. The output in Figure 10.1 is an all-forms search sorted by keyword from which a teacher has extracted promising formulaic pieces. Most concordance programs can perform these basic functions, with the difference between programs lying mainly in cost, ease of use, size of corpus treatable, whether exploiting a grammatically tagged or only "flat" corpus, and the level of sophistication in formatting the output (with Lextutor.ca at one extreme being free of cost and easy to use with smallish, flat corpora and the minor highlighting shown in Figure 10.1, and SketchEngine.co.uk, Kilgarriff, 2004, at the other providing colour-coded comparison insights from enormous corpora, either flat or tagged – but requiring some training to use and a user fee).

With whatever degree of sophistication, however, does the argument for corpus technology as a learning resource for FSs not extend beyond just "doing no harm"? In fact, the empirical case for substantial learning from corpus work is quite strong. In a meta-analysis of data-driven learning (DDL) approaches to language learning (involving the use of a corpus as a learning resource), Boulton and Cobb (2016) found that lexicogrammar (the category that FSs fall under)

Concordance extract for family RIDE
User selected output: 10 selected from 54 available in Corpus=corpus_graded_2k.txt on lextutor.ca/conc/

```
014.  was happy to agree that Brat would ride one of her horses, Chevron. Be
019.  er. After the boat there was a bus ride which took us past brightly pa
020.  rses. She hires her own jockeys to ride these horses in the Grand Nati
025.  I am a very keen swimmer and horse-rider. My hobby is collecting the p
026.  the 36-year-old Philadelphia woman rides a bike. She has become a fost
045.  Murdoch felt cross and tired as he rode home on his bicycle from schoo
046.  he bag into his bicycle basket and rode off towards Jericho. Yes you h
047.  r on his black horse, and with him rode a beautiful lady, her black cu
052.  riage, on a beautiful black horse, rode Quintus, the new Commander of
053.  free. He jumped onto the horse and rode fast toward one chariot. The d
```

FIGURE 10.1 Doing no harm with corpus data.

comprised 49 of the total 64 studies in their cull, with an average effect size of 1.54 standard deviations for within-groups studies (pre-post designs) and .75 standard deviations for between-groups studies (experimental and control groups designs). These effect sizes are "very large" and "large" by the field-specific standards of applied linguistics (Plonsky & Oswald, 2014, p. 889). Examples from this collection of FS studies with strong results include Chan and Liou (2005), Chen (2011), Daskalovska (2014), Huang (2014), Liou et al. (2006), and Sun and Wang (2003).

These DDL studies are quite diverse in the tasks they set for learners and the type of corpus presentation they adopt, but Chan and Liou (2005) provides a good example of the approach. The researchers had Chinese EFL learners use a bilingual Chinese-English parallel sentence concordance (sentences in English on one side and in Chinese on the other for a given search pattern) to fill gaps in a sentence which typically involved a pair of strongly associating verb-noun pairs ("The man tried to ____ fire to his neighbor's house with gas", with *set* the missing item). Learners were given strategies for discovering the missing word in the concordances (search for instances of *fire* sorted by the word on the left, which should pull out several instances of *set* in a corpus of any size) but pretty much left to work independently and develop personal search strategies. The overall effect size for this study was 2.41 standard deviations, compared to traditional ways of doing this same learning. This pedagogical sequence resembles many in the DDL tradition, namely a worksheet of some kind to be completed, usually collaboratively, through consulting and generalizing from corpus data.

Rationale for using corpus data to learn formulae

Even if corpus data was the source of linguists' discovery of the extent of formulaicity in language in the 1980s, and has similarly shown itself to help learners in the DDL studies to become aware of this generally and learn some patterns specifically, we still do not have any reason to believe that concordance work is uniquely positioned to help learners with FL. The unique value of corpus data in learning about formulae arises from the fact that formulae are "difficult" in the first place.

The difficulty of FL is notorious and well documented. (Since the learning of formulae is the topic of Chapter 8 in this volume, only aspects relevant to computation are discussed here.) As early as Bahns and Eldaw (1993), formulaicity in general, and collocation in particular, has consistently been described as the final aspect of language to be even partially mastered by language learners. The problem seems to be that formulae, whether for receptive or productive language use, are learned at a glacial pace through massive exposure that is relatively unmediated by intention or cognition, compared to comparatively straightforward word meanings or grammar patterns (Ellis, 1994). In language reception, Martinez and Murphy (2011) found that even when learners knew all the words in a reading

passage *qua* single words, they still had poor comprehension for its overall meaning when the passage contained idiomatic formulae (where individual words do not add up to a predictable meaning, like *beat around the bush*). In language production, the result of this glacial learning rate for FL is that learners with a strong grip on grammar and extensive single-word lexicons are often able to produce utterances that strike native speakers as "odd" or "foreign", which Kjellmer (1991) attributed to the fact that "in these learners' production, the building material is individual bricks [words] rather than prefabricated sections [lexicalised phrases]" (p. 124).

But is it really worth learners' trouble to get FL right? Apart from the avoidance of linguistic *faux-pas*, there is the larger matter that it is probable the smooth, automatic handling of FL is basic to the memory processing requirements of using a second or any language. Pawley and Syder (1983), at the beginning of the formulaic era and arguably its instigators, speculated that language processing would be an impossibly complex task if every word of an utterance had to be handled separately (as learners show they are doing when they ask, "Do you want to marry me?"). These researchers proposed instead that "native-like fluency" depends on large parts of language processing consisting of low-cost handling of formulaic patterns *qua* chunked, single items, with working memory thereby left free to handle a relatively small number of truly novel, unpredictable constructions.

The fact that recurring formulae were also problematic for professional linguists prior to the corpus era should give learners some cheer, in that the source of the problem was the same in both cases, and the linguists have solved it – or at least solved it on the level of awareness if not on the level of detail. What linguists discovered is that while formulaicity is pervasive in language, particular formulae sadly are not. Apart from a small number of extremely frequent formulae (as identified by Shin & Nation, 2008, or Martinez & Schmitt, 2012), the vast majority, though known to native speakers, are rare. A common joke at learners' expense is the misuse of the expression "to pull somebody's leg" (meaning to tease, which ends up as "pull somebody's legs", etc.). Why should this be so hard to learn? It is simple lack of exposure. Even in the TenTen corpus ("enTenTen13", comprising 22,728,686,012 word tokens, as analyzed by SketchEngine set to count all lemma variants and one intervening word), there are just 5,654 instances involving "pull" + my/your/his/etc. + "leg", or one instance per 5 million words, and about 15% of these are literals which do not involve the teasing idea.

The TenTen corpus probably represents something in the order of all the words a native speaker would hear or read in a lifetime. One humorous estimate of the words we speak in a lifetime is 860,341,500 (Brandreth, 1980), or about 1/26th of the TenTen, so if we hear or read 26 times as many as we produce, this estimate is not implausible for native speakers. For learners, a more realistic sample of the language they might encounter can be found in the purpose-built pedagogical corpus of 14 million words of basic English, including graded stories and informal speech in UK and US variants, compiled by Nation (2012) for use in the higher frequency levels of his *Range* software. In this corpus, there are 16 instances of

FIGURE 10.2 Years' worth of exposure assembled in a moment.

leg-pulling, including all legal family variants and sequences ("having my leg pulled" etc.), of which only nine involve the "teasing" metaphor, or one instance per nearly 2 million words. In other words, a learner who reads a million words a year might never encounter it. The individual items are of course far more numerous (2,372 for "pull', 1,503 for "leg"). So how would "pull somebody's leg" ever be observed, let alone learned, except by luck? It will happen only if a teacher knows it is a reasonably important thing to know at a certain stage of learning and uses some means like a concordance to bring it together (as shown in Figure 10.2).

And further, leg-pulling is a colourful idiomatic formula which is presumably easier to become aware of and learn than the huge number of far less striking literal formulae (e.g., riding bikes, not driving them), where computer search is if anything more needed to replace or supplement natural observation.

It is the thesis of this chapter that massive exposure to formula information is required for complete language learning, and that only computationally assembled data can provide this exposure. Concordance output can assemble more formula information, more effectively, than any other pedagogy is capable of. Making it pedagogically interesting is, of course, another story, to be dealt with in the following sections. But first, is a concordancing approach a truly necessary component of any pedagogy of FL?

Low-technology alternatives

It has been argued that the reading of texts in general, and graded readers in particular, can also be a source of formula acquisition, which takes place in a more contextually meaningful way than is provided by dissociated concordance lines (which in addition are not always simple to interpret, Figure 10.2). In an incidental acquisition study, Webb, Newton, and Chang (2013) found that if learners read while listening for roughly one hour to a graded story, of between 4,000 and 7,000 words within their existing vocabulary knowledge, seeded with different frequencies of the same formulaic expressions (verb-object collocations like *face*

facts and *blow nose*), then with 15 occurrences they learned to recognize appropriate matches from a multiple-choice selection in 75% of the 18 available test items, with productive knowledge only slightly less. In other words, formula acquisition can occur incidentally through reading and is sensitive to frequency (see discussion in Chapter 8).

But while this is an interesting result in principle, it is also highly limited as a pedagogy for formula acquisition. Learners will typically not be reading and listening at the same time (a particularly propitious arrangement for many types of vocabulary growth; Horst, Cobb, & Meara, 1998). They will not be reading texts with all vocabulary known other than the FSs. They will not meet the same FSs as many as 15 times in one hour of reading, though they will meet far more than 18 distinct FSs, of which many will also be unknown. They will not be given the base form of formulae to identify in a test (test items required only recognition of *face facts* or *blow nose*) but instead left to work these out for themselves over time from items with morphological variation and intervening items (*facing hard facts* and *blowing big noses*).

Would anything like this study's learning conditions be replicated in unseeded graded readers? The original text used in Webb et al. (2013), Oxford Bookworms' version of *New Yorkers* (from Henry, Hedge, & Bassett, 2000), when run through the "Clusters/N-Gram" feature of Anthony's (2014) *AntConc* Concordancer, yields surprisingly few formula learning opportunities. The story contains only 11 two to five word sequences that are repeated 15 times, and none of these are verb-object units, or any sort of independently meaningful unit. In other words, extensive reading might be a source in principle of learning FSs, if the learning opportunities were present, except that they will probably not be, and Webb et al. do not discuss the number of hours of unmodified graded readers that would be required for a comparable rate of acquisition.

For the demonstration of a frequency effect, and an attempt to find formula acquisition within a pleasurable context of story-reading, and an implementation of Boers et al.'s (2017) counsel that learners be exposed only to intact formula, Webb et al.'s (2013) study is commendable. Yet it still leaves us rather far from a practical solution to the FSs acquisition problem. The proposition of this chapter is that some form of corpus searching is at present the only complete and even provisionally proven pedagogy for making language learners aware of and proficient in the handling of FSs. The question is what form this corpus searching will take.

Contextualized corpus work

The worksheet and concordancer approach typical of DDL as described previously in the discussion of Chan and Liou (2005; namely worksheet for completion with corpus information leading to measurably raised awareness) is just one possible type of corpus work. The limits of this type of pedagogy are not hard

to imagine. The learning takes place in a very narrow semantic context, as may be reflected in the less impressive results at delayed post-test in Chan and Liou (as was also found across DDL studies generally by Boulton & Cobb's, 2017, meta-analysis). This possibly reflects the general truth of cognitive psychology (e.g., Anderson, 2015) that memories are not strong for low-meaning inputs such as might characterize disjointed worksheet questions or concordance lines. This explains the motivation to search for formula learning in more meaningful contexts such as reading graded stories, as Webb et al. (2013) do in their study. But concordancing work can be imagined, which is less decontextualized, with concordancing embedded within a more CALL-like environment, which typically keeps records, includes game elements, incorporates considerations of motivation, context of learning, etc.

The benefits of this idea are largely speculative to this point, however, since CALL developers and enterprises have not focused significantly on formula learning. In a review paper on CALL and the teaching of FL, Nesselhauf and Tshichold (2002) found this topic "largely neglected", and since none of the 68 Google Scholar citations to this paper up to November 2017 is a comparable treatment of the topic, their verdict appears to remain correct. One reason may be the ongoing migration of CALL vocabulary work to the small screen (*DuoLingo*, *Free Rice*, or the many other flashcard apps enumerated at https://en.wikipedia.org/wiki/List_of_flashcard_software), which for space constraints do not emphasize lexical information beyond the single word. Indeed, the present writer, as consultant to a gaming software project (reported in Cobb & Horst, 2012), witnessed first-hand the difficulty of evolving the single-word version of the Nintendo game *My Word Coach* to a v.2 incorporating formulaic information (the project was abandoned). For these reasons, much of the interesting work being done on CALL and modified concordancing approaches to integrating formulae within lexis takes place on the periphery of the CALL universe, on teacher-developer websites like Lextutor (www.lextutor.ca).

Two elements of the basic concordancing experience are modifiable within a CALL context: the nature of the corpus and the learning context.

Modifying the corpus

Concordance searches in applied linguistics are normally performed on full corpora which claim to represent a language as a whole, such as the British National Corpus (BNC; Oxford University Computing Services, 2001) or Corpus of Contemporary American English (COCA; Davies, 2008). Using such corpora, an applied linguist or teacher can determine something about a language that is relevant to learning it, such as that a particular formulaic expression like "pull someone's leg" is probably not frequent enough to be learned incidentally. But for pedagogical purposes, the corpus involved in a formula learning activity need not represent the language as a whole, but might instead reflect the purposes of

a learner or a teacher more than a linguist or applied linguist. Cobb (1999) had learners search for lexical information in a 50,000-word corpus assembled from the same learners' own study materials (e.g., course books, language lab assignments, classroom tests, and worksheets) such that some or most of the concordance lines in any given activity were probably familiar, especially when expanded to paragraph size. Interestingly, vocabulary acquired through the use of this corpus had not begun to dwindle by delayed post-test, as compared to some of the DDL studies mentioned earlier.

Modifying the learning context

Even if the language of a corpus is familiar, asking learners to use it to answer questions they do not currently have is nonetheless a decontextualized exercise. This is not the case however if the question comes from learners themselves. For example, learners are seeking corpus information in a meaningful context when they click on a word or expression in a text to see a concordance output for it, as shown in Figure 10.3, where concordance lines serve as a type of gloss. This technology has been provisionally validated for single-word learning by Lee, Warschauer, and Lee (2017). Users of such a tool are almost certainly seeking mainly single-word information in the concordance lines, but what they seek is not necessarily the only thing they get. Notice that in Figure 10.3 the concordances are set up by the developer to provide a secondary focus on any repeated sequences there may be in the output by left-of-keyword sorting. Thus, the most likely

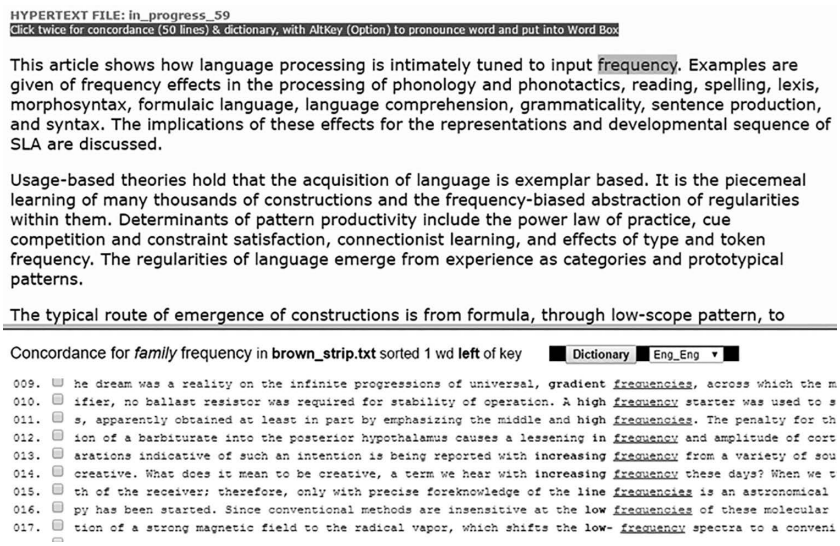

FIGURE 10.3 Sneaking formula insights into lexical search.

Text is from Ellis (2002); corpus is Brown (Kucera & Francis, 1971); routine is http://lextutor.ca/hyp/1/.

formulae are delivered along with the lexical information. Formula information is "sneaked in" to the single-word search for "frequency" ("high frequency", "increasing frequency", and "low frequency"). Whether learners notice this extra information has not been formally investigated.

Other examples of sneaked-in formula information on the Lextutor website are *List_Learn* (http://lextutor.ca/list_learn/), where learners build their own comprehensive flashcard glossaries by generating concordances from a word list to engage in a choose-the-definition quiz (shown in Figure 10.4, where "cease to" and particularly "cease to be" and "cease to exist" have been serendipitously generated); and *Concord_Writer* (www.lextutor.ca/cgi-bin/conc/write/), in which a writer's own emerging text is fully linked to a concordancer, such that "possible next words" can be generated from any of several corpora.

The learning context can also include what teachers do. Still in the context of a reading activity, a teacher can use a corpus tool to determine which repeated strings in a text they should draw learners' attention to. A teacher who runs a text through AntConc's or Lextutor's *N-Gram* routines can learn which strings are recurring, a feature which normally lies below the radar of even native-speaker awareness. Then, once such strings have been identified, these can form the content of concordance-based worksheets or quizzes tied to the text under study, on paper or online. The paper exercise shown in Figure 10.5 is made using Lextutor's concordance "gap" routine, where the keyword is replaced by a gap. The text in this case is about driving, and the search word is *car* with any of *drive*, *driver*, *truck*, or *ride* in the context. The main learning affordance here would be in a

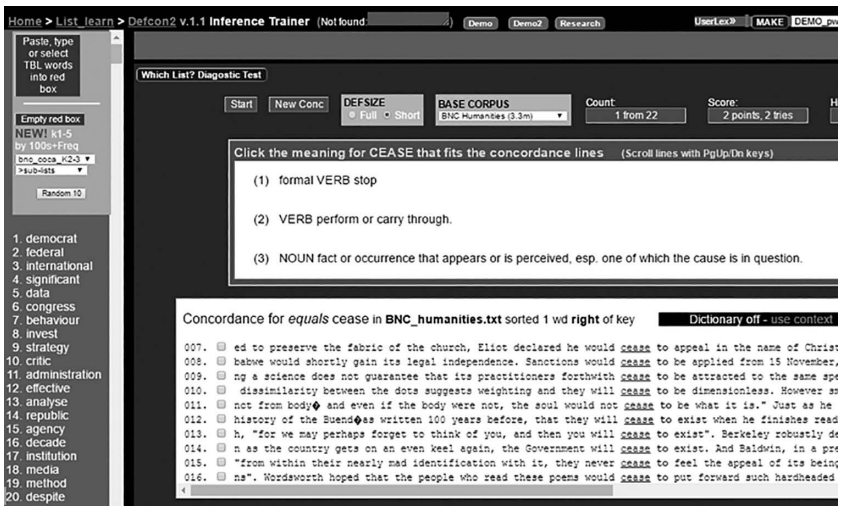

FIGURE 10.4 Combined word and formula trainer.

Defcon2 is from www.lextutor.ca/cgi-bin/list_learn/defcon2.pl; word lists are from Nation's (2012) BNC-COCA lists; dictionary is drawn from the *Concise Oxford Dictionary*; Corpus is *BNCs'* humanities subcorpus.

Instructions: Write in the word that fits all the gaps

001. it is a matter of fact that Smith cannot drive a ____. There is nothing to suggest that the brain ca
002. status when I am driven up in front of work in a ____ driven by my wife, who is only a woman. Even t
004. , North Providence, with injuries suffered when a ____ he was driving struck a utility pole on Woonas
007. nside of the door-frame on the driver's side of a ____. She called softly, 'Barney'. He looked in her
008. of appealing to them for help. Perhaps they had a ____ or truck and would drive him into town. Then h
013. For those who need or want and can afford another ____, buying one and driving it on the grand tour,
017. you insure five or more rigs. This means either a ____ or a truck. Discounts run up to 2% of cost. Us
020. avaughn Huntley is accused of driving the getaway ____ used in a robbery of the Woodyard Bros.' Groce
022. transported, may think of the engine driving his ____ as 'a mystical beast under the hood'. The Ital

FIGURE 10.5 CALL-concordance worksheet contextualized for post-reading.

From Brown corpus; made by routine at http://lextutor.ca/conc/eng.

teacher-led discussion of the roll of the words associated with *car* (*drive*, *driver*, *driving*, etc.) in determining the correct answer. A point to address here is that asking learners to supply a missing piece of a formula may look like an instance of the bad pedagogy discovered by Boers et al. (2017) and discussed prior, except that it targets and reviews an expression recently encountered.

A similar activity could also be put into the hands of learners through an online CALL-Concordancing activity. The activity shown in Figure 10.6 comes from a text with a strong presence of non-overlapping *make* and *do* verbs. Learners are asked to choose the word that will fill all the gaps in each concordance, and then repeat the exercise with a new randomization of concordance lines from the same or a different corpus. A paper task (like the one shown in Figure 10.5) can be simply created by the teacher to follow the online work, employing a new randomization, thus affording an opportunity for re-use of inputs and for transfer. Again, teacher involvement in discussion and feedback would help highlight the formulaic and collocational information that determines the correct answers.

Another learner-meaningful context for a CALL-Concordance activity is when the concordance is offered as a response to learners' writing errors. In the event of a writing error in an online submission, a teacher can use Lextutor's concordance input form (at http://lextutor.ca/conc/eng) to click together the pieces of a URL that generates a concordance bearing a more correct version of what the learner was trying to say, then copy-paste the URL into the learner's document to return for correction (the procedure is described with examples in Gaskell and Cobb, 2004). Here is an example of such a URL showing a CALL-concordance error feedback (Corpus is graded readers collection from OUP's Bookworm series; routine is Corpus Concordance English at http://lextutor.ca/conc/eng/):

www.lextutor.ca/cgi-bin/conc/wwwassocwords.pl?SearchStr=marry%20me&SearchType=equals&Corpus=corpus_graded_2k.txt&SortType=left&LineWidth=120&AssocWord=will&Fam_or_Word=word

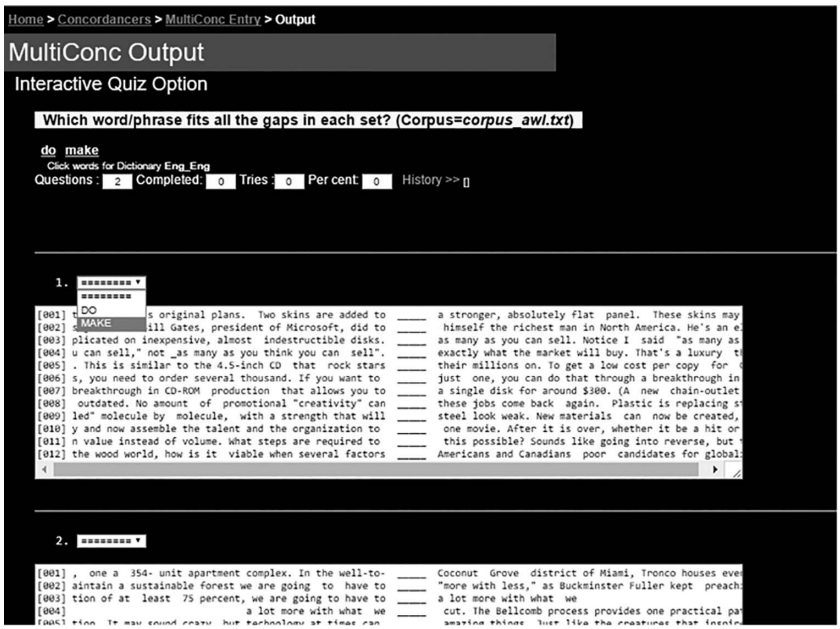

FIGURE 10.6 CALL-Concordancing repeatable Make-and-Do activity from a text.

Corpus is graded readers collection from OUP's Bookworm series; routine is MultiConcordance at http://lextutor.ca/conc/multi.

CALL-concordance feedback can be used with any type of productive error, but it is particularly apt for errors in FL. Figure 10.7 is taken from a tutorial routine employing typical formula errors that was developed to prepare learners in the Gaskell et al. study for corpus-based correction. Here the concordance for "marry me", from a corpus of graded readers, makes it reasonably clear that "Will you marry me?" is the standard formulation for this idea while "Do you want to marry me?" while not impossible is marked or non-standard. The learner considers the concordance information and then makes a correction in the corresponding space on the right. Not obvious in the screen-print is that the learner can easily reformulate the search, for example using "marry me" as the keyword and "want" as an associated word to the left. From this it will be obvious that while "want to marry me?" is used in certain contexts, "Do you want to marry me?" rarely appears as a direct question.

In summary, there are several ways of presenting concordancing information to learners in contexts they are normally motivated to attend to, with formulaic information as either the direct or indirect focus of the exercise.

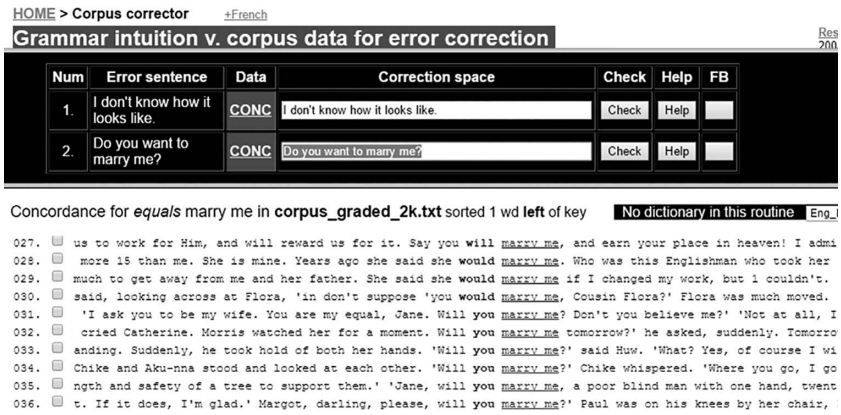

FIGURE 10.7 Corpus-based formula adjustment.

"Will you marry me?" is the archetypal formulaic sequence from Pawley and Syder (1983); Corpus corrector quiz is from www.lextutor.ca/conc/gram/.

CALL formula work other than concordancing

So far in this discussion it may have seemed that concordancing is the only possible way to have learners work with FL in CALL. But there are two other responses worth talking about: formulaic cloze passages and single-word work with formulaic spin-off.

Formulaic cloze passages

While CALL approaches to lexical development have traditionally focused on single words (Cobb & Horst, 2011), and indeed were often versions of the word-to-meaning or L1-to-L2 word flashcard concept (Nakata, 2011), various kinds of FSs are gradually being accommodated within the genre. Some of these involve drag-and-drop versions of the textbook activities found ineffective by Boers et al. (2017). The other major tendency in CALL vocabulary work that does not employ matching or word cards is the cloze passage, which presents words in a context but also has tended to work on the level of single words. There is no reason, however, that computer cloze passages cannot find and remove entire formulaic expressions for learners to return to their places in a text. Boers et al.'s counsel to have workers attend only to intact formulae is not thereby violated. In Figure 10.8, Lextutor's "Cloze Builder" displays the lyrics with accompanying sound file for David Bowie's song *Space Oddity* (1969). The teacher has chosen some intact (without intervenors) multiword units for removal, with varying degrees of success as to what constitutes a true formula, and the learner's task is to put them back in their original places. Resources to aid the learner in this task include the sung rendition

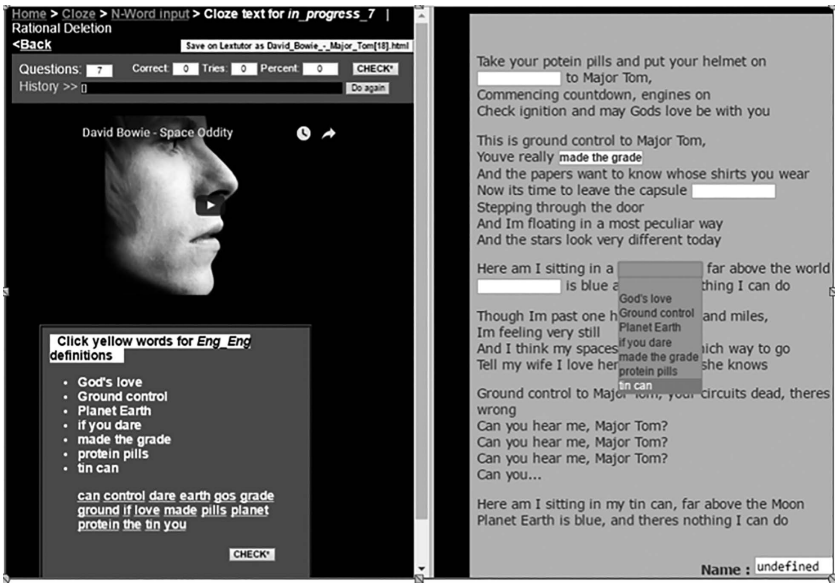

FIGURE 10.8 CALL cloze for sequences, not words.

From David Bowie's *Major Tom* (1969); cloze passage from http://lextutor.ca/cgi-bin/cloze/n/ ; video from www.youtube.com/watch?v=VrERLeFseDA; WordReference from http://mini.wordreference.com.

of the text, a listing of the items that have been removed, and a glossary for looking up their individual words in a mono- or bilingual dictionary (English to nine other languages). Apparently, no research has been done to determine the value of this type of activity for formula learning; the value will presumably be whatever it is for single-word cloze passages, possibly with additional awareness raising for the "togetherness" of some words and its accompanying prosody in the sound rendition.

Single-word CALL and formulaic sequences

On the single-word level, both the computer and its update, the mobile smart phone, have proven to be highly effective vocabulary teachers (Cobb, 1997; Cobb & Horst, 2011). However, CALL vocabulary approaches have focused almost entirely on single words, despite researchers' growing awareness of the existence and importance of recurring FSs. A single-word focus is probably inevitable and indeed acceptable in this context for two reasons. First is the relatively small number of single words that must be learned to get a basic grip on a language, namely 3,000 high-frequency families for 95% coverage in average texts (Schmitt, Cobb, Horst & Schmitt, 2017), compared to the truly stupendous number of FSs involving just those same few items (probably tens of thousands that,

though low in frequency, are nonetheless known to all native speakers). In other words, the practicality of direct teaching of all those FSs is of doubtful value. Who could remember them? Second is the fact that learning single words in any case appears to facilitate the subsequent learning of formulae in which they feature.

Some researchers have argued that given the prevalence of FSs, particularly within particular domains, single-word work is a waste of time (Hyland & Tse, 2007). This argument has been leveled particularly against the teaching of the 580 items of the corpus-based Academic Word List (Coxhead, 2000), whose items are shown to function differently within different formulaic expressions in different domains, leading to the possibility that this is how they should be presented and learned. Hyland et al's may be an extreme position, yet it is supported by a general lack of clarity within vocabulary research and instruction as to whether words or formulaic sequences should be taught, or when, in what proportions, and with what handover points.

There is some evidence, however, to suggest that assuming a dichotomy between single-word teaching and FS teaching is probably not useful. It now appears likely that single-word learning can be a precursor of multiword learning; that is to say, once a single-word form and basic meaning have been fixed in memory, then further learning of formulae involving that word is facilitated. And yet that the converse is not true: learning a formula does not necessarily facilitate learning the items that compose it. Wray's (2002) famous example is that few people can state what Rice Krispies are made of. There has been for some reason little investigation of this interesting question, apart from Bogaards (2001). He asked Dutch learners of French to learn idiomatic FSs (in which the meanings of the individual words did not add up to the intended overall meaning), half of whom had previously learned the individual words of the sequences and half had not. Those who had previously learned the words were significantly better able to both comprehend and retain the sequences.

This insight from French gradually made its way into English as a second language research. Nguyen and Webb (2017) looked for the predictors of L2 learners' learning of FSs, and found the strongest predictor to be the frequency of the node word of a sequence, this presumably reflecting roughly the number of times a learner would have seen this word *qua* single word. These researchers also found significant positive correlations between learners' knowledge of single words and their knowledge of FSs. In other words, learning words and learning sequences tends to go hand in hand, and there is little evidence of a word-formula dichotomy. Indeed, it is known to any language teacher that there are few learners with large single-word lexicons who do not also know large numbers of FSs, and vice versa.

This is not to say that the whole process from single-word to multiword learning needs only be left to nature to happen by itself, or that it happens most efficiently left to itself. Martinez and Murphy's (2011) subjects knew all the single words in a story passage rich in idiomatic formulae but nonetheless did not

comprehend the main idea of the passage. In other words, single-word learning had not (yet) transferred to formulaic knowledge, at least in the case of idiomatic formulae. This suggests that specific training in the interpretation of word groups is required, though it is not at present clear whether this should amount to awareness raising for formulae generally or focused work with particular formulae.

What does all this mean for CALL and the acquisition of formulaic sequences? Basically, that the single-word successes of CALL probably did not come at the expense of formulaic skill but rather facilitated it, long term, for the many learners who have used this software in the past 20 years – but also that more could probably be done to raise awareness of formula at the same time as working on single words. The CALL program that is needed is hence one that is legitimately focused mainly on single-word acquisition, but with some simultaneous attention to the groupings and formulae that these words are likely to enter ("sneaked in" in the idiom adopted above, but perhaps more explicitly).

Conclusions and future directions

This chapter has argued that corpus technology is key to exposing the extent and nature of formulaic language to second and foreign language learners. Without seeing "pull someone's leg" in a corpus, learners will never notice it, or if they do, have any way to evaluate its importance as a learning object. A carefully built and properly analyzed corpus can show both the extent of the formulaic phenomenon (raise awareness of it) as well as the usage characteristics of particular formula. Such a corpus can do this in an engaging manner without either breaking formulae apart (and running the risks elaborated by Boers et al., 2017) or misrepresenting their distribution characteristics (as Webb et al., 2013, have implicitly done). However, the chapter also argues, concordance work can probably be accomplished best in a pedagogical context where motivation, curriculum integration, and learner purpose are taken into account – that is, in a CALL context.

But as mentioned, however, Nesselhauf and Tshichold's (2002) conclusion was that formula instruction was "largely neglected in CALL", and this would appear to still be the case with two exceptions, concordance work in the DDL approach and a handful of CALL experiments that integrate concordancing within a tutorial context. In other words, despite the historical connection between text computing and FL, the connection has not been extensively exploited instructionally. A number of operational ideas for integrating formula work within ongoing CALL vocabulary work have been shown in this chapter, and some of it has received preliminary empirical validation (Boulton & Cobb, 2017, for lexicogrammar by DDL generally; Gaskell & Cobb, 2004, for concordance error feedback; Lee et al., 2017, for concordance glossing). The strong result for DDL in lexicogrammar will presumably transfer positively to effectively designed CALL-concordancing, with its greater attention to learner variables and pedagogy.

An issue humming in the background of this chapter is the unresolved question whether formulae can be tackled in detail in L2 instruction (particular formulae highlighted and practiced), or only generally in the form of awareness raising (learners made aware they should pay attention to how words go together). Both sides of this question have their proponents. Simpson-Vlatch and Ellis (2010) built and then crunched a corpus of academic lectures for specific to-be-learned formulae for that setting; Thornbury (2002) argues that for general English, the number of formulae and quasi-formulae is so vast that only an awareness-raising approach can be effective. In the opinion of the preseent writer, corpus work appears to be effective whichever answer turns out to be correct.

The progress of this work will to some extent await resolution of some other issues in learning research and technology development. Questions that will have a bearing on this and which to some extent fall out of concerns raised in this chapter are the following:

- Should formula work involve mainly awareness raising or mainly teaching of particular formulae, or does this depend on the learning purpose?
- Should FSs be taught before, after, or along with the single words they are composed of?
- Can formulae knowledge and skill be incorporated within the small screen and short learning time that have typified recent CALL success in vocabulary acquisition, or is a different technology or paradigm needed?
- If the vocabulary money is now riding on the small screen, and the small screen is inherently unsuited to formulaic information, where will this new paradigm come from?
- Is there any need for learners to perform corpus analysis themselves, or can they simply be shown the insights of experts who have done so?

The ideal way forward in this area can be indicated by the behaviours and strategies of learners themselves, the best of whom already use the Web as their prime resource for information about formulae and which words go together. But it should be possible for applied linguistics as a profession to do better than just let learners roam the Web for their language insights. We should be able to offer them learning technologies for formula work that incorporate specialized corpora, feedback, motivation to persist, opportunities to review – in short, pedagogy.

References

Anderson, J. (2015). *Cognitive psychology & its implications* (8th Ed.). New York, NY: Worth Publishers.

Anthony, L. (2014). *Antconc 3.4.4w*. Computer program. Retrieved October 20, 2017 from www.laurenceanthony.net/software/antconc/.

Bahns, J., & Eldaw, M. (1993). Should we teach EFL students collocations? *System*, *21*(1), 101–114.

Brandreth, G. (1980). *The joy of lex: How to have fun with 860,341,500 words*. New York, NY: Morrow.

Boers, F., Dang, T., & Strong, B. (2017). Comparing the effectiveness of phrase-focused exercises: A partial replication of Boers, Demecheleer, Coxhead, and Webb (2014). *Language Teaching Research*, *21*(3), 362–380.

Bogaards, P. (2001). Lexical units and the learning of foreign language vocabulary. *Studies in Second Language Acquisition*, *23*(3), 321–343.

Boulton, A., & Cobb, T. (2016). Corpus use in language learning: A meta-analysis. *Language Learning*, *65*(2), 1–46.

Chan, T. P., & Liou, H. C. (2005). Effects of web-based concordancing instruction on EFL students' learning of verb – Noun collocations. *Computer Assisted Language Learning*, *18*(3), 231–251.

Chen, H.-J. (2011). Developing and evaluating a web-based collocation retrieval tool for EFL students and teachers. *Computer Assisted Language Learning*, *24*(1), 59–76.

Cobb, T. (1997). Is there any measurable learning from hands-on concordancing? *System*, *25*(3), 301–315.

Cobb, T. (1999). Applying constructivism: A test for the learner-as-scientist. *Educational Technology Research & Development*, *47*(3), 15–31.

Cobb, T., & Horst, M. (2011). Does word coach coach words? *CALICO Journal*, *28*(3), 639–661.

Cowie, A. P. (1998). *Phraseology: Theory, analysis, & applications*. Oxford: Oxford University Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238.

Daskalovska, N. (2014). Corpus-based versus traditional learning of collocations. *Computer Assisted Language Learning*, *28*(2), 130–144.

Davies, M. (2008). *The corpus of contemporary American English*. Provost, UT: Brigham Young University.

Ellis, N. (1994). Vocabulary acquisition: The implicit ins & outs of explicit cognitive mediation In N. Ellis (Ed.), *Implicit & explicit learning of languages* (pp. 211–282). London: Academic Press.

Ellis, N. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*, 143–188.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, *20*(1), 29–62.

Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System*, *32*(3), 301–319.

Henry, O., Hedge, T., & Bassett, J. (2000). *New Yorkers: Short stories* [Oxford Bookworms Library Stage 2]. Oxford: Oxford University Press.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, *11*(2), 207–233.

Huang, Z. (2014). The effects of paper-based DDL on the acquisition of lexico-grammatical patterns in L2 writing. *ReCALL*, *26*(2), 163–183.

Hyland, K., & Tse, P. (2007). Is there an 'Academic Vocabulary'? *TESOL Quarterly*, *41*(2), 235–253.

Kilgarriff, A. (2004). *The sketch engine*. Computer program. Retrieved October 20, 2017 from www.sketchengine.co.uk/.

Kjellmer, G. (1991). A mint of phrases. In Aijmer, K., & Altenberg, B. (Eds.), *English corpus linguistics* (pp. 111–127). London: Longman.

Kucera, W., & Francis, H. (1971). *The brown corpus of present-day edited American English*. Princeton: University Press.

Lee, H., Warschauer, M., & Lee, J. (2017). The effects of concordance-based electronic glosses on L2 vocabulary learning. *Language Learning & Technology*, *21*(2), 32–51.

Liou, H.-C., Chang, J. S., Chen, H. J., Lin, C.-C., Liaw, M.-L., Gao, Z. M., Jang, J.-S. R., Yeh, Y., Chuang, T. C., & You, G.-N. (2006). Corpora processing and computational scaffolding for an innovative web-based English learning environment. *CALICO Journal*, *24*(1), 77–95.

Martinez, R., & Murphy, V. (2011). Effect of frequency and idiomaticity on second language reading comprehension. *TESOL Quarterly*, *45*(2), 267–290.

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics*, *33*(3), 299–320.

Nakata, T. (2011). Computer-assisted second language vocabulary learning in a paired-associate paradigm: A critical examination of flashcard software. *Computer Assisted Language Learning*, *24*, 17–38.

Nation, P. (2012). *The BNC/COCA words family lists*. Retrieved November 8, 2017 from www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Information-on-the-BNC_COCA-word-family-lists.pdf.

Nesselhauf, N., & Tshichold, C. (2002). Collocations in CALL: An investigation of vocabulary-building software for EFL. *Computer Assisted Language Learning*, *15*(3), 251–279.

Nguyen, T., & Webb, S. (2017). Examining second language receptive knowledge of collocation and factors that affect learning. *Language Teaching Research*, *21*(3), 298–320.

Oxford University Computing Services. (2001). *The British National Corpus*, version 2 (BNC World). Distributed on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/.

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

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912.

Shin, D., & Nation, P. (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal*, *62*, 339–348.

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

Sinclair, J. (1991). *Corpus, concordance, collocation*. London: Oxford University Press.

Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use English? Replication of Van Zeeland & Schmitt (2012), Nation (2006), and Cobb (2007). *Language Teaching*, *50*(2), 212–226.

Sun, Y.-C., & Wang, L.-Y. (2003). Concordancers in the EFL classroom: Cognitive approaches and collocation difficulty. *Computer Assisted Language Learning*, *16*(1), 83–94.

Thornbury, S. (2002). *How to teach vocabulary*. Harlow: Longman.

Webb, S., Newton, J., & Chang, A. (2013). Incidental learning of collocation. *Language Learning*, *63*(1), 91–120.

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