

Corpus linguistics to tackle the learning and teaching of multiword units: Necessary but not sufficient

Gaëtanelle Gilquin
UCLouvain

Abstract

Aims and Objectives

The article aims to show that corpora have made it possible to expand the field of phraseology and gain a better understanding of multiword units, including in foreign or second language learning and teaching. At the same time, however, it argues that corpora cannot answer all questions related to multiword units in and for learners and it therefore pleads for the combination of corpora with other sources of information.

Approach

The article first describes the contributions of corpora to phraseology, emphasizing that several essential aspects of multiword units are not accessible to intuition and require the use of corpora. It then highlights the limitations of corpus linguistics when it comes to the study of multiword units in and for learners and shows how these limitations can be overcome by combining corpora with other sources of information.

Findings

Linguistic expertise, experimentation and pedagogical experience are presented as useful complements to the corpus study of multiword units in and for learners, allowing for the investigation of additional aspects such as the functional classification of multiword units, their mental representations or their teachability. The article mentions different ways of combining these sources of information and underlines some of the challenges involved in combining them.

Originality and Significance

While multiword units have often been approached from a single perspective, using one source of information, this article argues for methodological pluralism and triangulation, claiming that multiword units in and for learners can be approached more comprehensively through the combination of several sources of information.

1. Introduction

Multiword units, which can be broadly defined as sequences of words that go together and display some (semantic, pragmatic, etc.) unity, have been approached from different perspectives and investigated by means of different methods and data. One approach that has been particularly popular is that of corpus linguistics. By relying on electronic databases of authentic language, corpus linguists have offered many new insights into the use of multiword units. Several important aspects of multiword units would not be amenable to scientific study without recourse to corpora, which makes corpus linguistics necessary for a comprehensive analysis of multiword units. Yet, it is not true that corpus linguistics is sufficient to tackle

multiword units, as some issues may require additional sources of information. This article shows why and how such sources of information should be used in combination with corpora.

The focus is on foreign- or second-language (L2) learning and teaching and hence on multiword units in learners (language learning) and for learners (language teaching). Corpora can help gain a better understanding of how learners acquire and use multiword units: how they misuse, overuse or underuse certain multiword units; how they may transfer features of multiword units from their native language (L1) to the target language; how they may mix languages in the production of multiword units and so on. Corpora can also provide insights that are relevant to the teaching of multiword units, for example, to determine what multiword units should be included in the curriculum, help teachers prepare pedagogical material or let students make their own findings about language. However, this still leaves many questions about multiword units in and for learners unanswered, such as the following: What linguistic features are necessary for a sequence of words produced by learners to qualify as a multiword unit? How are multiword units processed by learners? How do native speakers react to erroneous multiword units? What is the most efficient way of teaching multiword units? For such questions to be answered, other sources of information are necessary, as will be argued in this article.

The article starts by showing how corpus linguistics has transformed and expanded the field of phraseology (Section 2) and helped overcome the limitations of intuition (Section 3). It also shows, however, that corpus linguistics has its own limitations when it comes to the study of multiword units (Section 4). This pleads for the combination of corpus linguistics with other sources of information, three of which are described and illustrated in Section 5, namely linguistic expertise, experimentation and pedagogical experience. While the combination of several sources of information may present some challenges (Section 6), it also makes it possible to approach multiword units more comprehensively, gaining knowledge that corpus linguistics alone could not gain. Section 7 concludes the article.

2. Corpora and multiword units

Two main approaches to the study of multiword units can be distinguished (see, e.g., Granger & Paquot, 2008; Nesselhauf, 2004). The first one, usually referred to as the phraseological approach, is the more traditional one. It is a top-down approach based on linguistic criteria, such as compositionality or fixedness, and paying particular attention to very idiomatic units like *kick the bucket* or *It's raining cats and dogs*. The other approach, called the distributional or frequency-based approach, is a bottom-up approach which relies on lexical co-occurrence and tends to focus on highly frequent units (e.g., *make progress*, *of the*).

The distributional or frequency-based approach is a product of corpus linguistics, which has offered tools and techniques to extract recurrent word combinations in a corpus-driven fashion. Thanks to this approach, corpus linguistics has made two significant contributions to the study of multiword units. First, it has highlighted the diversity of multiword units, beyond idioms, proverbs, and other formulae, to also include phenomena such as collocations (words that tend to co-occur within a certain distance of each other, e.g., *raise* and *issue*), colligations (words that tend to occur with a specific grammatical category, e.g., *cause* and *to*-infinitives), n-grams/lexical bundles (recurrent sequences of contiguous words, e.g., *has suggested that*), collocational patterns/p-frames (recurrent sequences with one or more variable slots, e.g., *on the * of*) and collostructions (constructions that attract certain words in a specific slot, e.g.,

[SUBJ *give* OBJ OBJ]). Second, corpus linguistics has demonstrated the pervasiveness of multiword units in language. Erman and Warren (2000) claim that some 59% of spoken English discourse and 52% of written English discourse consist of phraseological language. Through these contributions, corpus linguistics has opened the door to multiword units that are less restricted, phraseologically speaking, but more pervasive in language. The impact of the distributional or frequency-based approach has been such that, as Granger and Paquot (2008: 29) note, it has “literally turned phraseology on its head”.

The demonstration of the pervasiveness of phraseology has led to the recognition of its importance in L2 learning and teaching (e.g., Ebeling & Hasselgård, 2015; Meunier & Granger, 2008). Not only can the use of (appropriate) phraseology help learners produce language that is typical of the target language, but it has also been shown to facilitate language production. By relying on ready-made chunks, learners are able to free up some of their attention and apply it to other aspects of language production, leading, for example, to more fluent speech (see Wood, 2010). Given the benefits of using multiword units, scholars have pleaded for their integration into the L2 curriculum. Szudarski (2023, p. 45) argues that “information on phrasal vocabulary needs to feature systematically in teaching syllabi and materials, with insights from corpora enriching and informing pedagogical decisions of teachers and textbook writers”, thus also underlining the role that corpora have to play in this undertaking.

Pedagogical emphasis on multiword units is all the more important since, as pointed out by Granger and Bestgen (2014, p. 229), “phraseology is one of the aspects that unmistakably distinguishes native speakers of a language from L2 learners”, as corpus research has repeatedly proved. Learners often have “phraseological teddy bears”, that is, multiword units that they particularly like and use over and over again, like *on the other hand* for Norwegian learners (Hasselgård, 2019). Compared to native speakers of the language, they underuse certain multiword units, for example phrasal verbs among learners with a Romance language as their L1 (Gilquin, 2025). They also transfer some multiword units from their L1, by translating a multiword unit word for word, as illustrated by the use of *make an experience*, instead of *have an experience*, by German learners, who seem to be influenced by the German phrase *eine Erfahrungen machen* (Nesselhauf, 2005). They sometimes use stylistically inappropriate multiword units, for example, *of course* in academic essays (Gilquin & Paquot, 2008). Finally, learners sometimes memorize chunks from textbooks, such as *last but not least* among Chinese learners (Leedham & Cai, 2013), which can lead to the overuse of the chunk and sometimes to noncanonical variants in case the learner has not memorized the chunk properly.

3. The limitations of intuition to study multiword units

Despite the important contribution of corpus linguistics to phraseology, one may wonder whether corpora are required to study multiword units. In particular, intuition could be thought to be an equally good source of information about multiword units. This, however, does not appear to be the case, as the literature has emphasized the limitations of intuition (including native intuition) for the investigation of multiword units. Biber et al. (1996, p. 120) note that “[i]ntuitions regarding lexical associations are often unreliable and inaccurate”. We tend to have little awareness of the existence of multiword units. As Hunston (2002, p. 21) points out, “[i]t is difficult for the native speaker to be conscious of these combinations, and others like them, without corpus evidence”. We also tend to have poor intuitions about the frequency of multiword units. Stubbs (1995, p. 25) claims that native speakers “cannot give accurate

estimates of the frequency and distribution of different collocations”. In McGee (2009), teachers proved generally unable to identify the most frequent collocates of stimulus words. This may be linked to the difficulty of perceiving certain multiword units. In an experiment by Sosa and MacFarlane (2002), subjects took longer to recognize the presence of the word *of* in frequent collocations like *kind of* than in less frequent collocations like *piece of*. This suggests that our perception of multiword units and their frequency might be biased. Intuitions about learners’ use of multiword units do not fare better. This is the case even among specialists such as textbook writers, whose intuitions are often considered insufficient to report on learners’ knowledge of multiword units in the target language. This has led Möller (2020, p. 118) to state that “it is highly desirable that textbook writers move from intuitions about learner language to empirical evidence”.

Intuitions are arguably even less reliable among learners. While certain scholars acknowledge that native speakers may have some intuitions about multiword units, they do not think that learners are capable of such intuitions. This can be illustrated by this quote from Meunier (2012, p. 112): “L2 language learners find formulaicity particularly challenging as it is impossible for them to use the innate native intuition usually associated with formulaic language use”. Similarly, Granger (1998, p. 152) notes that learners have “an underdeveloped sense of salience and of what constitutes a significant collocation”. Her claim is based on an experiment in which learners were asked to mark in a list of adverb-adjective combinations those combinations that they believed were particularly salient. It turned out that learners hardly marked any of them and missed many of the significant collocations (e.g., *readily available*).

According to Stubbs (1995, p. 24), attested data like those found in corpora are required given the unreliability of intuitions as a source of evidence about multiword units. Corpora are ideally suited to extract different types of multiword units. Those identified through lexical co-occurrence, in particular, can be extracted automatically from corpora, relying on a range of statistical measures to determine the most significant multiword units (mutual information, t-score, log-likelihood, etc.). Corpora can also supply information, both quantitative and qualitative, about multiword units as they occur in naturalistic, authentic language. Provided adequate corpora are available, multiword units can be explored in any languages, language varieties (e.g., Canadian English, English produced by Italian-speaking learners) and registers (e.g., academic writing, classroom talk) or any combination thereof (e.g., comparison of multiword units in native and learner Spanish).

While other methods can serve to investigate certain aspects of multiword units, as will be seen later, the identification of frequent or strongly associated multiword units as well as their description in their natural environment would not be possible, at least not with such a high degree of precision and reliability, without access to corpora. This shows that corpora are required to study some essential aspects of multiword units.

4. The limitations of corpora to study multiword units

The fact that corpora are required to investigate essential properties of multiword units does not mean that they are sufficient for the study of multiword units as a whole. In fact, corpora present some limitations which make them insufficient for a comprehensive coverage of multiword units.

First, while multiword units extracted from corpora can be examined for features such as compositionality or degree of fixedness, which are central to the more traditional approach to multiword units (see above), corpora in and of themselves do not provide information about such features. Thus, corpus techniques do not distinguish between compositional and noncompositional word sequences or, for that matter, between (frequent) free combinations and phraseological units. While they can distinguish between types of multiword units that are formally different (e.g., bigrams vs. trigrams), they cannot distinguish between types of multiword units that differ semantically or pragmatically (e.g., proverbs vs. commonplaces). This is one of the challenges faced by linguists studying multiword units (see Schmid, 2023) to which corpora cannot provide a direct answer.

Secondly, it is a well-known limitation of corpora that they are necessarily finite (Stefanowitsch, 2020, p. 5-7). Some corpora may be too small to include certain existing multiword units or the criteria adopted, such as a minimum frequency threshold, may not make it possible to retrieve them. While mega-corpora (including web-based ones) may partially mitigate this issue, no corpora, however large, will include the infinite possibilities of (phraseological) language production. Many of the most popular idioms have actually been shown to be extremely rare in corpora (e.g., Grant, 2005; Moon, 1998). The idiom *spill the beans*, for instance, has a very low frequency of between 0.02 and 0.39 occurrences per million words in the corpora examined by Moon (1998).

A third limitation is that corpus data do not give access to mental representations of multiword units – or if they do, this is only in a very indirect way, for example, with frequency being taken as a possible sign of cognitive entrenchment (cf. Schmid, 2010). As pointed out by Hyland (2012, p. 152), the frequency-based retrieval of multiword units (such as the lexical bundles that Hyland is interested in) “allows us to say little about how such sequences are psycholinguistically processed”.

Corpora also have specific weaknesses for the study of multiword units in learner language, some of which are related to the limitations outlined above. First, in relation to the finiteness of corpora, those that are made up of data produced by learners (learner corpora) are usually quite small in comparison with corpora of native language. They may therefore not offer enough evidence for certain types of multiword units that tend to be less common, for example, long n-grams (see Ebeling & Hasselgard, 2015).

Then, because corpora do not give access to mental representations, we cannot investigate learners’ competence on their basis (Gilquin, 2021a). Learner corpora only reflect learners’ production (their performance), which is not necessarily the same as their actual knowledge (their competence). Learners may know other multiword units than those that they produce, or they may only know one meaning of a multiword unit that they use, but such insights are not recoverable from corpora.

It should also be borne in mind that a fully automatic approach to multiword units in learner corpora may produce misleading results. Spelling errors, for instance, which can be quite frequent in learner language, may prevent the automatic retrieval of recurrent word sequences, simply because the words are spelt differently and hence are not aggregated, as illustrated by *private correspondence* and *private correspondance* (Bestgen & Granger, 2014, p. 32). Similarly, noncanonical variants of multiword units, such as *some words about* or *couple of words about* (Mauranen, 2010, p. 20-21), might be missed because the concordancer will treat them separately from the standard variant *a few words about*. In addition, learner corpora do

not say anything about the accuracy of the multiword units retrieved from them, unless the learner corpora have been error-tagged (cf. Huang, 2015). A large amount of lexical co-occurrence in a learner corpus is therefore not always indicative of a high degree of phraseological proficiency (see, e.g., Tan & Römer, 2022).

Finally, as far as teaching applications are concerned, corpus techniques as such do not distinguish between multiword units that are pedagogically relevant to learners and those that are not. It has thus been suggested that many of the n-grams extracted from corpora may not be useful to teach. Simpson-Vlach and Ellis (2010, p. 493) note, for example, that high-frequency n-grams such as *and this is* or *this is the* are “neither terribly functional nor pedagogically compelling”.

These limitations show that corpora cannot be sufficient to investigate multiword units in learners and for learners. They also raise the question of what else may be needed to study multiword units for the purposes of language learning and teaching. In the next section, three complementary sources of information are considered, namely linguistic expertise, experimentation and pedagogical experience. It is shown how combining these with corpus data can improve our understanding of multiword units.

5. Complementary sources of information about multiword units

5.1. Linguistic expertise

While corpus techniques can identify highly recurrent and strongly associated word combinations, linguistic expertise is still needed. This may seem obvious, but the automatic extraction of many types of multiword units from corpora could sometimes give the impression that the computer does all the work and that linguistic expertise is no longer required.

Linguistic expertise is necessary to determine the linguistic status of word combinations that are not formally distinguished, such as proverbs and commonplaces. It is also necessary to disambiguate corpus-extracted occurrences. Searching for phrasal verbs on the basis of the particle, for example, will produce irrelevant hits that need to be weeded out, a task that requires a good understanding of what a phrasal verb is (cf. Gilquin, 2025). Next, linguistic expertise is needed to code the retrieved data for further analysis and classify them into categories. For instance, a study may involve deciding whether multiword units are used with a figurative or nonfigurative meaning (e.g., Nacey, 2013) or categorizing lexical bundles functionally (e.g., Chen & Baker, 2010). In addition, linguistic expertise is required to interpret the results, checking for cases of possible L1 transfer, topic influence and so on (e.g., Juknevičienė & Grabowski, 2018). More generally, linguistic expertise makes it possible to ground the study of multiword units in a theoretical framework (e.g., second language acquisition, usage-based linguistics) which can clarify some of the issues at hand and help reach a higher degree of generalization and abstraction.

Two examples can serve to illustrate the important role of linguistic expertise in the study of multiword units in and for learners. The first one is taken from Nesselhauf’s (2005) study of verb-noun collocations in learner English. Her definition of collocation is of the more traditional type (see above), but her source of information is a learner corpus from which she retrieves all verb-noun combinations. In her book, she describes at great length how collocations were distinguished from other types of word combinations, pointing out that

“applying the definition [of collocation] to the actual data proved to be one of the greatest challenges of the study” (Nesselhauf, 2005, p. 54). In order to meet this challenge, she followed a whole procedure to establish the combinability of verbs (i.e., whether the verbs are restricted in the nouns that they can combine with), a decisive criterion for her to determine whether the verb-noun combinations were collocations or not. For example, *want* can combine with almost any noun, whereas *commit* can only combine with a small set of nouns, with which it therefore forms collocations (Nesselhauf, 2005, p. 30). Nesselhauf (2005) relied on her own linguistic expertise, but also that of lexicographers (by consulting monolingual dictionaries) and native speakers (through acceptability judgments). In addition, she checked certain verb-noun combinations in a corpus of native English to see the extent of their combinability.

The second example is a study by Biber et al. (2004) on lexical bundles in corpora representing classroom talk and textbooks. Although lexical bundles were extracted automatically, their analysis involved linguistic expertise to develop a functional taxonomy for classifying the lexical bundles. The authors examined the typical uses of each bundle in the corpus data and grouped together the bundles fulfilling similar functions. They then identified the discourse functions of each group of bundles, coming up with labels such as “topic elaboration” (e.g., *has to do with*, *was going to say*) or “specification of attributes” (e.g., *in the form of*, *the nature of the*). In addition to their own linguistic expertise, Biber et al. (2004) relied on that of other linguists, through reference to the previous literature. They also needed linguistic expertise to classify the data, a task that turned out to be challenging due, among other things, to the multifunctionality of some multiword units.

5.2. Experimentation

Experimentation has been shown to be an ideal complement to corpus approaches (see Gilquin, 2021a; Gilquin & Gries, 2009). This applies to the study of multiword units in and for learners too. Corpus data reveal what multiword units learners produce, get right or wrong, overuse or underuse, but experimentation makes it possible to go further and consider types and aspects of multiword units that are difficult or impossible to investigate in corpora.

Experimental data, first, allow for the study of multiword units that are rare in learner corpora. Thus, Bardovi-Harlig and Stringer (2017) use discourse completion tasks to collect instances of pragmatic formulae (e.g., *Sorry I'm late*), which are typically hard to find in learner corpora. Through experimentation, one can also manipulate the context, for example, comparing politeness formulae used to ask for a small favour or a big one. In corpora, on the other hand, one is dependent on the contexts that actually occur in the data.

In addition, experimentation can help determine what multiword units learners know (receptively or productively), that is, their competence, as opposed to learner corpora, which merely give access to what language they produce, that is, their performance. In Gilquin (2007), for example, a corpus analysis of collocations with *make* produced by French-speaking learners of English reveals an error rate of only 7%. By contrast, experimental data from fill-in and acceptability judgment tests display error rates of 43% and 51%, respectively. This demonstrates that learners' knowledge of *make* collocations is much more limited than suggested by their error rate in unconstrained language production. Further analysis of the data provides possible explanations for this discrepancy, including learners' preference for congruent collocations in their language production, that is, collocations that have a direct equivalent in their L1 (e.g. *make progress* = *faire des progrès* in French; *make an effort* = *faire*

un effort in French) and are therefore less likely to lead to errors. When learners are forced to make decisions about noncongruent collocations in the fill-in and acceptability judgment tests, on the other hand, they often make errors. This points to a possible phenomenon of avoidance in unconstrained learner language production and an overreliance on familiar collocations, which may hide learners' actual deficiencies.

Like competence, mental representations are only accessible through experimentation, which can give information about how multiword units are stored and processed in the learner's mind. Kim and Kim (2012), for example, extracted two-word combinations from a corpus of American English, distinguishing between various levels of frequency. These combinations were embedded in sentences which native speakers and learners of English were asked to read word by word during a self-paced reading task. The results show that more frequent combinations tend to be read faster, which suggests that they are stored holistically in the mind and retrieved as wholes. Interestingly, this processing advantage only holds for highly frequent combinations among learners, whereas among native speakers it holds for all combinations except the very infrequent ones.

Thanks to experimental data, it is also possible to check how native speakers react to noncanonical multiword units found in learner corpora: whether they consider these units acceptable or not (e.g., Geeraert et al., 2017), how they process them (e.g., Millar, 2011) and so on. Millar (2011), for instance, demonstrates on the basis of a self-paced reading task that nonstandard collocations found in learner corpora (e.g., *marriage life*) are read significantly more slowly by native speakers of English than their standard equivalents (e.g., *married life*). This leads him to the conclusion that collocation errors in learner language place “an increased processing burden on native speaker addressees” and “have the *potential* to hinder communication” (Millar, 2011, p. 142; emphasis original).

Finally, from a more pedagogical point of view, one can conduct experiments to test the feasibility and/or efficiency of phraseology-focused instruction (e.g. De Knop & Mollica, 2016; Gilquin, 2021b). Using a within-subjects experimental design, Gilquin (2021b) seeks to test the efficiency of teaching multiword units by giving learners access to corpus instances, a pedagogical method known as data-driven learning (DDL; Gilquin & Granger 2022). The so-called *way* construction (see Perek, 2018), as illustrated by *worm one's way into* or *work one's way through*, was the target of one of the teaching interventions. In order to evaluate the efficiency of the DDL approach, sentence production tests were carried out to compare learners' knowledge of the construction before and after the intervention. It turns out that learners produced more (correct) instances of the *way* construction right after the teaching intervention than before. Interestingly, a week or two after the session, the number of sentences produced and the proportion of correct sentences dropped, but it went up again later, which suggests that it might have taken learners some time to truly benefit from the long-term effects of the DDL session.

5.3. Pedagogical experience

A third source of information which can come as a complement to corpus approaches is pedagogical experience. It is crucial if one is interested in multiword units for learners, that is, teaching applications. Corpora show the multiword units typical of learner vs. native language and even those typical of teacher talk (e.g., Nicaise, 2021). However, pedagogical experience is needed to make decisions about what implications, if any, this should have for teaching. As

Granger and Meunier (2008, p. 250) put it, “it is not because something is frequent in the language that [it] is necessarily worth teaching”.

Pedagogical experience is necessary to decide what multiword units are pedagogically relevant. Some scholars, for example, have claimed that it is the most frequent chunks that should be taught (e.g., Kennedy, 2008), while others have argued that high-frequency chunks can easily be picked up from language input and that instruction should therefore focus on medium-frequency chunks (Boers et al., 2010). One may also wonder whether it is pedagogically relevant to teach certain n-grams, especially those that are structurally incomplete, such as *at the end of the*, or whether, for pedagogical purposes, it would help to “complete” such n-grams, as suggested by Cappelle and Grabar (2016), teaching *at the end of the N* instead of *at the end of the*.

In addition, pedagogical experience can help decide what is teachable about multiword units. Corpora provide a wealth of (quantitative and qualitative) information about all sorts of multiword units, in different registers, varieties and so on, but it is not feasible to teach students all of these minute details. The *Collins COBUILD Grammar Patterns* volumes (Francis et al., 1996, 1998) and the *Valency Dictionary of English* (Herbst et al., 2004) are examples of corpus-based resources that are very rich and thorough but that could presumably not be used as such in most classrooms because they would likely be overwhelming for students.

While corpora can uncover multiword units that are potentially relevant to students as well as aspects of multiword units that are potentially teachable, and while corpora can be used by students themselves to learn multiword units (for example, through DDL), it takes pedagogical skills to implement all of this. Educators can decide how best to teach multiword units, possibly according to students’ specific profiles (e.g., students with knowledge of other L2s, students with special educational needs) and according to the teaching context (e.g., Content and Language Integrated Learning or English-Medium Instruction settings). They can also come up with teaching activities and exercises that are ideally suited to practise multiword units.

N. Ellis and Simpson-Vlach’s (2009) investigation of n-grams is an example of a study that takes pedagogical experience into account. A number of n-grams were extracted from corpora of academic speech and writing, corresponding to different lengths, different frequency bands and different association (mutual information) bands. English language instructors were asked to judge the instructional validity of these n-grams. More precisely, they had to answer three questions (N. Ellis & Simpson-Vlach, 2009, p. 65), using a scale from 1 (“I disagree”) to 5 (“I agree”):

1. Do you think the phrase constitutes “a formulaic expression, or fixed phrase, or chunk”?
2. Do you think the phrase has “a cohesive meaning or function, as a phrase”?
3. Do you think the phrase is “worth teaching, as a bona fide phrase or expression”?

The results reveal that n-grams like *in other words* obtained much higher scores among the language instructors than n-grams like *that the only*. While in this case the instructional validity correlates with corpus-derived measures (the frequency of the n-grams and, especially, mutual information) and the same information could therefore be derived from corpora, it is still noteworthy that N. Ellis and Simpson-Vlach (2009) included language instructors as a source of information. As Urzúa (2015, p. 99) rightly points out, “[t]he language teacher is an often neglected figure in learner corpora projects”. Letting teachers’ voices be heard in research on

multiword units for learners ensures that the selection of corpus-derived findings is pedagogically sound.

6. Triangulating information about multiword units

The idea of relying on several sources of information to tackle a linguistic phenomenon is hardly new or original. Methodological pluralism, triangulation and the search for converging evidence have become very popular in linguistics. More than 20 years ago, Čermák (2002, p. 270) already noted that

corpora must not be offered as a cure-all. There will always be other resources to be tapped whenever complex information is needed. It is not true that one can find everything in corpora only.

Talking about learner language more specifically, R. Ellis and Barkhuizen (2005, p. 49) pointed out that “no one method will provide an entirely valid picture of what a learner knows or thinks”. A number of publications have argued for the combination of multiple methods in language studies, including Gilquin and Gries (2009), Riazi and Candlin (2014) and Schönefeld (2011). In their literature review, Gilquin and Gries (2009, p. 16) found that a large proportion of the studies combining corpus and experimental data focused on phraseological phenomena, which suggests that multimethod approaches may be particularly fruitful for the analysis of multiword units.

Corpora can be combined with any number of additional sources of information, either incrementally or iteratively. As the starting point of an incremental approach, multiword units could be extracted automatically from a corpus (e.g., sequences of verbs and particles to retrieve phrasal verbs). Relying on linguistic expertise, one could then weed out the corpus-extracted sequences that are not actually phrasal verbs, select among the phrasal verbs those that fulfil certain criteria (e.g., figurative phrasal verbs) and use these as a basis to design an experiment (e.g., a self-paced reading task with figurative phrasal verbs). Finally, the results of the experiment could be combined with teaching experience to develop useful pedagogical material (e.g., a gap-fill exercise for figurative phrasal verbs that turned out to be processed holistically and that teachers consider pedagogically relevant).

In order to take full advantage of all the strengths of each source of information, these sources could be made to interact in multiple and recurrent ways, forming a more cyclical or iterative process. Corpora could be used at different steps of the process: to analyse naturally-occurring instances of certain multiword units, to find experimental stimuli, to collect information about multiword units produced in an experiment (e.g., formulae collected through a discourse completion task), to provide educators with material for the development of pedagogical resources for the teaching of multiword units and to allow students to do their own research about multiword units in a DDL activity. Teaching experience could be used right from the start to formulate a research question about multiword units that is pedagogically relevant, in the corpus study to select for further analysis certain multiword units that are useful to the target students, in the experiment to design test items with multiword units that are understandable to students with a certain proficiency level and, of course, in the pedagogical application of the findings. Linguistic expertise could be used all along the way to inform and refine the study, ground it in a solid theoretical framework, interpret its results and guide teachers' decisions.

Despite the relevance of combining corpus data with other sources of information to study multiword units in and for learners, there are some potential challenges to bear in mind. One of them has to do with the possibility of diverging evidence. In case not all sources of information point in the same direction, one should decide what source of information (if any) should be given priority. If corpus and experimental methods indicate that a certain multiword unit is frequent and holistically stored, for example, but teachers say that it is pedagogically irrelevant or not teachable, the teacher's advice should presumably be given priority, and the multiword unit should not be taught. If, on the other hand, a teacher thinks that multiword units should be taught in a certain way but experimentation reveals that this teaching method is not efficient, the experimental results should take precedence. Consider, however, a situation in which a sequence of words is frequent in corpora but does not appear to be processed holistically. It may not be easy, in such a case, to decide what source of evidence should take priority and whether the sequence should be taught as a multiword unit or as individual items. Generally speaking, scholars and teachers should be trusted to make the best decisions based on the empirical evidence that is available to them.

Another major challenge is the interdisciplinarity that this kind of multimethod approach involves. It requires combining different skills as an individual (having linguistic expertise, corpus literacy, teacher experience, the capacity to set up a rigorous experimental design, etc.) or collaborating with other experts (psycholinguists, language teachers, etc.). The latter, however, can be made difficult by the fact that everybody comes from different disciplines, each with its own traditions and theories. It is therefore crucial that the experts should start by discussing basic concepts (including the concept of multiword unit) and how they understand them, explore the different methodologies that exist in their disciplines and agree on important aspects such as the type of multiword units to investigate, how to identify them, how to combine the different methods, what counts as evidence for what, how much weight to give to each source of information and so on. Keeping an open mind and testing different procedures are key to ensuring a fruitful collaboration, such as the one described in Römer (2016) between a corpus linguist, a computational linguist, a genre expert, a cognitive linguist and a psycholinguist.

7. Conclusion

It has been shown that corpora provide information about multiword units that is not accessible to intuition and could not be gathered otherwise. The frequency-based approach, in particular, requires the use of corpus data. Without corpora, a whole section of phraseology would therefore escape our attention. We would also be unable to produce (reliable) quantitative insights into multiword units and describe their use in authentic contexts. As evidenced by the pre-corpus view of phraseology, this is likely to place undue emphasis on certain phraseological phenomena, namely those that are striking but not necessarily common in naturally-occurring language, such as idiomatic expressions. In the field of L2 learning and teaching, this biased view of phraseology may be particularly detrimental, as it could deprive learners of core knowledge which might improve their phraseological proficiency, but also their language performance as a whole (most notably their fluency).

From that perspective, one may argue that corpora are necessary to investigate multiword units, because a non-corpus-based approach to phraseology runs the risk of omitting important types of multiword units (e.g., lexical bundles) and important aspects of their study (e.g., quantification, authenticity). This does not mean that all studies of multiword units need to rely

on corpus data. It is perfectly valid to carry out a study of multiword units as defined in a more traditional (i.e., non-frequency-based) manner and use, say, linguistic expertise and experimental data to develop a better understanding of these multiword units. The field of phraseology as a whole, however, cannot dispense with the knowledge that is gained from the analysis of corpora, including in and for learners. This seems to be confirmed by the many experimental studies on multiword units that use corpora to prepare stimuli (e.g., Kim & Kim, 2012; Pulido, 2023).

Yet, it has also been suggested that studies relying exclusively on corpora may miss key aspects of multiword units, some of which are relevant to L2 learning and teaching (e.g., learners' knowledge of multiword units, efficiency of methods for teaching multiword units). This pleads for the combination of corpus data with other sources of information (linguistic expertise, experimentation, pedagogical experience, etc.) to investigate multiword units, especially in and for learners. These sources of information can be combined incrementally or, better still, iteratively, so as to take full advantage of all of their respective strengths. Such a combination is not without its challenges (e.g., having to deal with diverging evidence or managing the inherent interdisciplinarity). However, it is worth the additional effort, since it allows for a more comprehensive approach to phraseology, one that should help advance the field but also, ultimately, benefit language learners.

References

- Bardovi-Harlig, K., & Stringer, D. (2017). Unconventional expressions: Productive syntax in the L2 acquisition of formulaic language. *Second Language Research*, 33(1), 61–90. <https://doi.org/10.1177/0267658316641725>
- Bestgen, Y. & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, 26, 28–41. <https://doi.org/10.1016/j.jslw.2014.09.004>
- Biber, D., Conrad, S., & Cortes, V. (2004). *If you look at ...: Lexical bundles in university teaching and textbooks*. *Applied Linguistics*, 25(3), 371–405. <https://doi.org/10.1093/applin/25.3.371>
- Biber, D., Conrad, S., & Reppen, R. (1996). Corpus-based investigations of language use. *Annual Review of Applied Linguistics*, 16, 115–135. <https://doi.org/10.1017/S0267190500001471>
- Boers, F., Deconinck, J., & Lindstromberg, S. (2010). Choosing motivated chunks for teaching. In S. De Knop, F. Boers, & T. De Rycker (Eds.), *Fostering language teaching efficiency through cognitive linguistics* (pp. 239–256). Berlin: De Gruyter. <https://doi.org/10.1515/9783110245837.239>
- Cappelle, B., & Grabar, N. (2016). Towards an n-grammar of English. In S. De Knop & G. Gilquin (Eds.), *Applied construction grammar* (pp. 271–302). Berlin: De Gruyter. <https://doi.org/10.1515/9783110458268-011>
- Čermák, F. 2002. Today's corpus linguistics. Some open questions. *International Journal of Corpus Linguistics*, 7(2), 265–282. <https://doi.org/10.1075/ijcl.7.2.06cer>
- Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, 14(2), 30–49. <http://dx.doi.org/10.125/44213>
- De Knop, S. & Mollica, F. (2016). A construction-based analysis of German ditransitive phraseologisms for language pedagogy. In S. De Knop, & G. Gilquin (Eds.), *Applied construction grammar* (pp. 53–88). Berlin: De Gruyter. <https://doi.org/10.1515/9783110458268-004>

- Ebeling, S., & Hasselgård, H. (2015). Learner corpora and phraseology. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 207–230). Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139649414.010>
- Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.
- Ellis, N., & Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory*, 5(1), 61–78. <https://doi.org/10.1515/CLLT.2009.003>
- Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, 20(1), 29–62. <https://doi.org/10.1515/text.1.2000.20.1.29>
- Francis, G., Hunston, S., & Manning, E. 1996. *Collins COBUILD grammar patterns 1: Verbs*. New York: HarperCollins Publishers.
- Francis, G., Hunston, S., & Manning, E. 1998. *Collins COBUILD grammar patterns 2: Nouns and adjectives*. New York: HarperCollins Publishers.
- Geeraert, K., Newman, J., & Baayen, R. H. (2017). Idiom variation: Experimental data and a blueprint of a computational model. *Topics in Cognitive Science*, 9(3), 653–669. <https://doi.org/10.1111/tops.12263>
- 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. <https://doi.org/10.1515/zaa.2007.55.3.273>
- Gilquin, G. (2021a). Combining learner corpora and experimental methods. In N. Tracy-Ventura, & M. Paquot (Eds.), *The Routledge handbook of second language acquisition and corpora* (pp. 133–144). New York: Routledge.
- Gilquin, G. (2021b). Using corpora to foster L2 construction learning: A data-driven learning experiment. *International Journal of Applied Linguistics*, 31(2), 229–247. <https://doi.org/10.1111/ijal.12317>
- Gilquin, G. (2025). Second and foreign language learners: The effect of language exposure on the use of English phrasal verbs. *International Journal of Bilingualism*, 29(2), 456–473. <https://doi.org/10.1177/13670069231169123>
- Gilquin, G., & Granger, S. (2022). Using data-driven learning in language teaching. In A. O’Keeffe, & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics. Second edition* (pp. 430–442). London: Routledge.
- Gilquin, G., & Gries, S. Th. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, 5(1), 1–26. <https://doi.org/10.1515/CLLT.2009.001>
- Gilquin, G., & Paquot, M. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction*, 1(1), 41–61. <https://doi.org/10.1075/etc.1.1.05gil>
- Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 145–160). Oxford: Oxford University Press. <https://doi.org/10.1093/oso/9780198294252.003.007>
- Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching*, 52(3), 229–252. <https://doi.org/10.1515/iral-2014-0011>
- Granger, S., & Meunier, F. (2008). Phraseology in language learning and teaching: Where to from here? In F. Meunier, & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 247–252). Amsterdam: John Benjamins. <https://doi.org/10.1075/z.138.19gra>

- Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger, & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 27–49). Amsterdam: John Benjamins. <https://doi.org/10.1075/z.139.07gra>
- Grant, L. E. (2005). Frequency of ‘core idioms’ in the British National Corpus (BNC). *International Journal of Corpus Linguistics*, 10(4), 429–451. <https://doi.org/10.1075/ijcl.10.4.03gra>
- Hasselgård, H. (2019). Phraseological teddy bears: Frequent lexical bundles in academic writing by Norwegian learners and native speakers of English. In V. Wiegand, & M. Mahlberg (Eds.), *Corpus linguistics, context and culture* (pp. 339–362). Berlin: de Gruyter. <https://doi.org/10.1515/9783110489071-013>
- Herbst, Th., Heath, D., Roe, I. F., & Götz, D. (2004). *A valency dictionary of English: A corpus-based analysis of the complementation patterns of English verbs, nouns and adjectives*. Berlin: De Gruyter. <https://doi.org/10.1515/9783110892581>
- Huang, K. (2015). More does not mean better: Frequency and accuracy analysis of lexical bundles in Chinese EFL learners’ essay writing. *System*, 53, 13–23. <https://doi.org/10.1016/j.system.2015.06.011>
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics*, 32, 150–169. <https://doi.org/10.1017/S0267190512000037>
- Juknevičienė, R., & Grabowski, Ł. (2018). Comparing formulaicity of learner writing through phrase-frames: A corpus-driven study of Lithuanian and Polish EFL student writing. *Research in Language*, 16(3), 303–323. <https://doi.org/10.2478/rela-2018-0016>
- Kennedy, G. (2008). Phraseology and language pedagogy: Semantic preference associated with English verbs in the British National Corpus. In F. Meunier, & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 21–41). Amsterdam: John Benjamins. <https://doi.org/10.1075/z.138.05ken>
- Kim, S. H., & Kim, J. H. (2012). Frequency effects in L2 multiword unit processing: Evidence from self-paced reading. *TESOL Quarterly*, 46(4), 831–841. <https://doi.org/10.1002/tesq.66>
- Leedham, M., & Cai, G. (2013). *Besides ... on the other hand*: Using a corpus approach to explore the influence of teaching materials on Chinese students’ use of linking adverbials. *Journal of Second Language Writing*, 22(4), 374–389. <https://doi.org/10.1016/j.jslw.2013.07.002>
- Mauranen, A. (2010). Features of English as a lingua franca in academia. *Helsinki English Studies*, 6, 6–28.
- McGee, I. (2009). Adjective-noun collocations in elicited and corpus data: Similarities, differences, and the whys and wherefores. *Corpus Linguistics and Linguistic Theory*, 5(1), 79–103. <https://doi.org/10.1515/CLLT.2009.004>
- Meunier, F. (2012). Formulaic language and language teaching. *Annual Review of Applied Linguistics*, 32, 111–129. <https://doi.org/10.1017/S0267190512000128>
- Meunier, F., & Granger, S. (Eds.). (2008). *Phraseology in foreign language learning and teaching*. Amsterdam: John Benjamins.
- Millar, N. (2011). The processing of malformed formulaic language. *Applied Linguistics*, 32(2), 129–148. <https://doi.org/10.1093/applin/amq035>
- Möller, V. (2020). From pedagogical input to learner output: Conditionals in EFL and CLIL teaching materials and learner language. *Pedagogical Linguistics*, 1(2), 95–124. <http://doi.org/10.1075/pl.00001.mol>
- Moon, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach*. Oxford: Clarendon Press.
- Nacey, S. (2013). *Metaphors in learner English*. Amsterdam: John Benjamins.

- Nesselhauf, N. (2004). What are collocations? In D. J. Allerton, N. Nesselhauf, & P. Skandera (Eds.), *Phraseological units: Basic concepts and their application* (pp. 1–21). Basel: Schwabe Verlag.
- Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.
- Nicaise, E. (2021). *Native and non-native teacher talk in the EFL classroom: A corpus-informed study*. London: Routledge.
- Perek, F. (2018). Recent change in the productivity and schematicity of the *way*-construction: A distributional semantic analysis. *Corpus Linguistics and Linguistic Theory*, 14(1), 65–97. <https://doi.org/10.1515/cllt-2016-0014>
- Pulido, M. F. (2023). Processing conventional and non-conventional multiword units: evidence of similarity-based generalisation from judgements and brain potentials. *Language, Cognition and Neuroscience*, 38(5), 651–671, <https://doi.org/10.1080/23273798.2022.2157028>
- Riazi, A. M., & Candlin, C. N. (2014). Mixed-methods research in language teaching and learning: Opportunities, issues and challenges. *Language Teaching*, 47(2), 135–173. <http://doi.org/10.1017/S0261444813000505>
- Römer, U. (2016). Teaming up and mixing methods: Collaborative and cross-disciplinary work in corpus research on phraseology. *Corpora*, 11(1): 113–129. <https://doi.org/10.3366/cor.2016.0087>
- Schmid, H. J. (2010). Does frequency in text instantiate entrenchment in the cognitive system? In D. Glynn & K. Fischer (Eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches* (pp. 101–133). Berlin: De Gruyter.
- Schmid, H. J. (2023). Can linguistic theory be relevant? Multi-word units and language teaching. Keynote speech at the International Workshop on Multi-Word Units in Multilingual Learners, München, 30 November – 1 December 2023.
- Schönefeld, D. (Ed.) (2011). *Converging evidence: Methodological and theoretical issues for linguistic research*. Amsterdam: John Benjamins.
- Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31(4), 487–512, <https://doi.org/10.1093/applin/amp058>
- Sosa, A. V., & MacFarlane, J. (2002). Evidence for frequency-based constituents in the mental lexicon: collocations involving the word *of*. *Brain and Language*, 83(2), 227–236. [https://doi.org/10.1016/S0093-934X\(02\)00032-9](https://doi.org/10.1016/S0093-934X(02)00032-9)
- Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Berlin: Language Science Press.
- Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language*, 2(1), 23–55. <https://doi.org/10.1075/fol.2.1.03stu>
- Szudarski, P. (2023). Corpora and teaching vocabulary and phraseology. In R. R. Jablonkai, & E. Csomay (Eds.), *The Routledge handbook of corpora and English language teaching and learning* (pp. 41–55). London: Routledge.
- Tan, Y., & Römer, U. (2022). Using phrase-frames to trace the language development of L1 Chinese learners of English. *System*, 108, 102844. <https://doi.org/10.1016/j.system.2022.102844>
- Urzúa, A. (2015). Corpora, context, and language teachers: Teacher involvement in a local learner corpus project. In V. Cortes, & E. Csomay (Eds.), *Corpus-based research in applied linguistics: Studies in honor of Doug Biber* (pp. 99–122). Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.66.05urz>
- Wood, D. (2010). *Formulaic language and second language speech fluency: Background, evidence and classroom applications*. London: Continuum.