The Routledge Handbook of Vocabulary Studies

Stuart Webb

Resources for Learning Multiword Items

Publication details
Fanny Meunier
Published online on: 13 Aug 2019

How to cite: Fanny Meunier. 13 Aug 2019, Resources for Learning Multiword Items from: The Routledge Handbook of Vocabulary Studies Routledge
Accessed on: 05 Dec 2023

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Introduction

Some of the preceding sections in the volume have addressed multiword items (MWI) from various angles: defining (types of) MWI (see Wood, this volume); factors affecting the learning of MWI (see Boers, this volume); learning single-word items vs. multiword items (see Pellicer-Sánchez, this volume); and processing single-word and multiword items (see Conklin, this volume). The present chapter specifically targets resources for learning MWI. However, before discussing such resources in more detail, I will briefly comment on the two views of “learning” that have been adopted to frame this section: the learning-for-use perspective (Edelson, 2001) and the Technological Pedagogical Content Knowledge (TPACK) model (Mishra & Koehler, 2006). It should be stressed that I have deliberately chosen to focus on learning in instructed settings as I consider teachers/instructors as key actors in presenting learners with useful resources to foster the acquisition and use of MWI. I would also like to add that while MWI is the preferred term throughout the volume, I may at times use the terms multiword units, phrasemes (phraseology), or formulae (formulaic language) as synonyms.

I have adopted the learning for use perspective, initially developed by Edelson in 2001, as it supports “the design of learning activities that achieve both content and process learning” (2001, p. 356). While Edelson’s focus was on technology-supported science teaching, his approach can easily be applied to language learning and teaching. Learning-for-use rejects memorization, which often leads to “inert knowledge” that learners/users can often not retrieve when they most need it, and supports (the design of) learning activities aiming at both content and process learning. Inert knowledge was first referred to in 1929 by Whitehead, a late 19th to early 20th-century English mathematician and philosopher. He was one of the first proponents of the process philosophy school and explained that inert knowledge is information/facts/knowledge that someone can express but not use in real situations. Typical examples of inert knowledge in language learning would be recently studied vocabulary items available for a specific test but no longer accessible in a real communicative situation, or the correct explicit wording of a grammar rule which is later not applied in a context where it
would clearly be required. The four key tenets of the learning-for-use model (Edelson, 2001, p. 357) are:

- Learning takes place through the construction and modification of knowledge structures.
- Knowledge construction is a goal-directed process that is guided by a combination of conscious and unconscious understanding of goals.
- The circumstances in which knowledge is constructed and subsequently used determine its accessibility for future use.
- Knowledge must be constructed in a form that supports use before it can be applied.

Such tenets very nicely tie in with current communicative and usage-based approaches to language learning in general, and to the learning of MWI in particular. Usage-based analyses of language (be it in corpus linguistics, cognitive linguistics, or psycholinguistics) have provided ample evidence of the highly formulaic nature of language. As stated by Ellis, Römer, and Brook O’Donnell (2016, p. 45) a number of psychological factors “conspire in the acquisition and use of any linguistic construction”, both in first (L1) language acquisition or in the acquisition of additional languages (L2, L3, etc.). MWI are ubiquitous linguistic constructions and have been shown to be subjected to a number of interconnected determinants of learning that include (1) frequency effects; (2) categorization, meaning, and prototypes; (3) contingency (the association of forms and meanings); (4) salience; and (5) various forms of learning (implicit vs. explicit). See for instance Harris, Murphy, and Rehder (2008), Smith (2014), Ellis (2017) and Siyanova-Chanturia and Pellicer-Sánchez (2019) for lengthier explanations on such determinants of learning. Native speakers implicitly (unconsciously) acquire and use MWI thanks to the preceding cited determinants of learning. Non-native speakers, in contrast, while also being sensitive to statistical patterns of use, are subject to cross-linguistic influence interfering in the process. Adopting a foreign language learning perspective, Meunier (2012, p. 111) suggests:

- Language as a whole is highly conventional.
- Even when alternative combinations exist only one is usually preferred.
- Rules and formalisms are often of little help in the acquisition of such combinations, and may be considered to increase teaching and learning difficulty.

Frequency information, prototypes, and form-meaning mappings are less directly accessible to non-native speakers than to native speakers. This view, also shared by Liu, Liu, Yu, Li, and Wen (2014, p. 682), led the authors to state that “EFL learners heavily depend on technology for learning authentic English”, precisely because technology can help learners access, among other things, frequency information or form-meaning mappings.

The second view of language that has been adopted in this part of the volume is the Technological Pedagogical Content Knowledge model. It is the second view on learning that will be used as a structuring device in this chapter. Mishra and Koehler’s (2006) and Koehler and Mishra’s (2008) Technological Pedagogical Content Knowledge model nicely complements the usage-based, learning-for-use model, and technological aspects mentioned in the previous paragraph. While the usage-based view concerns language as such and learning-for-use is more learner centered, the Technological Pedagogical Content Knowledge model can be said to be teacher oriented. As the use of technology can enhance foreign language teaching and learning, Liu et al. (2014, p. 683) argue that “EFL teachers need to ‘technologize’ their professional knowledge”. In that respect, Koehler and Mishra’s (2008) Technological
Pedagogical Content Knowledge model encompasses three main types of knowledge, as explained by Koehler et al. (2013, pp. 14–16) and summarized as follows:

- **Content knowledge (CK)** is the teachers’ knowledge about the subject matter to be learned or taught. In an L2 learning context, it is the knowledge of the target language itself.
- **Pedagogical knowledge (PK)** is teachers’ knowledge about the appropriate and up-to-date processes and practices or methods of teaching and learning (including the knowledge of overall educational purposes, values, and aims, the understanding of how students learn. For example, in L2 vocabulary learning this would involve knowledge of theoretical and applied perspectives on Second Language Acquisition, general classroom management skills, lesson planning capacities, and student assessment skills).
- **Technological knowledge (TK)**, constantly in flux and going beyond traditional computer literacy, includes an in-depth understanding of information technology which enables the teacher not only to apply technology productively but also to assess whether information technology can assist or impede the achievement of a learning goal.

These three types of knowledge can interact in dyadic or triadic ways. Technology and content knowledge (TCK) can for instance be strongly intertwined. Koehler et al. explain that “progress in fields as diverse as medicine, history, archeology, and physics have coincided with the development of new technologies that afford the representation and manipulation of data in new and fruitful ways” (2013, pp. 15–16). They list the discovery of X-rays in medicine or the technique of carbon-14 dating in archeology. Another typical dyadic interaction is that between technological and pedagogical Knowledge (TPK). This includes an understanding of “how teaching and learning can change when particular technologies are used in particular ways” (ibid). As will be explained in the next section, (learner) corpus research and its associated technologies and tools have been instrumental in L2 learning. They have directly contributed to technology and content knowledge by offering “a new way of thinking about language” (Leech, 1992, p. 106) but also to technology and pedagogical knowledge as the technological affordances of (learner) corpus research have made it possible to revisit the pedagogical approaches used in foreign language learning.

As for the all-encompassing triadic Technological Pedagogical Content Knowledge model, it is defined as “an understanding that emerges from interactions among content, pedagogy, and technology knowledge” (Koehler et al., 2013, p. 16). In the coming sections the focus will thus be on technological resources that can help language learners/users/teachers better understand SLA processes and access information such as frequency, form and meaning pairing, and “situated use” of MWI so as to enhance the construction and modification of language knowledge structures and foster subsequent use of MWI. Situated use (Norén & Linell, 2007, p. 390) entails that “(a) theory of meaning potentials assumes that parts of a word’s meaning are evoked, activated or materialized, foregrounded or backgrounded, in different ways in the different contexts, in which it is exploited”. Situated use is typically something that corpora can give plenty of access to, as will be explained in the next section.

**Critical Issues and Topics**

**Technical and Content Knowledge: Identifying MWI – (Learner) Corpus Research as a Window on Situated Use**

Viana, Zyngier, and Barnbrook (2011) explain that the partnership between the collection and analysis of corpora containing hundreds of millions of words enabled scholars “to expect
answers to questions that it would have been impracticable to ask a generation ago” (Crystal, 2003, p. 447, in Viana et al., 2011, p. 133). This “corpus revolution”, as Rundell and Stock (1992) call it, made it possible to describe “the way people actually use words when they communicate with one another” (Rundell, 2008, p. 23). The centrality of context, as postulated in situated language use (Norén & Linell, 2007, p. 390), has thus found a very concrete embodiment in corpus linguistics, and particularly so when it comes to MWI. In a terminological discussion of the notions of genres, registers, text types, domains, and styles, Lee (2001, p. 38) explains that “it is impossible to make many useful generalizations about ‘the English language’ or ‘general English’ since these are abstract constructions. Instead, it is far easier and theoretically more sound to talk about the language of different genres of text, or the language(s) used in different domains, or the different types of register available in a language, and so forth”. Corpora (be they native or learner corpora) and corpus tools have played a particularly significant role in “the increased focus on phraseological research” (Oksefjell Ebeling & Hasselgård, 2015, p. 207) and it is now possible to (semi-)automatically retrieve MWI in all sorts of text types and genres. Those MWI include lexical bundles (e.g., Biber, Conrad, & Cortes, 2004), n-grams (e.g., Stubbs, 2009), collocations (e.g., Sinclair, 1991), collostructions (e.g., Stefanowitsch & Gries, 2003), or constructions of various levels and types (Ellis et al., 2016, p. 17) (see also Wood in this volume for more details on the different types of MWI).

Leaving aside what Granger (2009b, p. 63) calls the “terminological chaos that besets theoretical phraseological studies”, it can be stated that corpus research has helped uncover target like MWI in native corpora while learner corpus research has helped “teachers identify the lexical, grammatical and discourse features that differentiate learners’ production from the targeted norm” (Granger, 2009a, p. 19). Adopting a formulaic approach to L2 learning and teaching has been shown to be relevant for three main reasons (Meunier, 2012, p. 112): (1) formulaicity is ubiquitous in language; (2) formulaic language use has been shown to be a marker of proficiency in an L2; and (3) studies have demonstrated that 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.

Illustrations: Concrete Impact of (Learner) Corpus Research on the Production of Learning Materials of Various Types

Some publishers are now using corpus data to inform textbook writing. Meunier (2012) gives concrete examples of corpus-informed textbooks and discusses the advantages and limitations of such materials, with a specific focus on MWI. Some textbooks now include sections such as collocations, metaphors, idioms, phrasal verbs, frequent lexical chunks or communicative phrasemes in conversational English. Mishan and Timmis (2015) devote a whole chapter of their Materials Development for TESOL volume to corpus-informed materials for teaching vocabulary and grammar. They argue for the benefit of using corpora as resources. Farr (2015) also devotes a whole section of her book to corpus-based materials, discussing concepts, findings, and literature on the topic. Farr (2015) also proposes tasks to explore corpora (see for instance Task 3.9 on p. 41) and offers a list of corpus-based reference books, corpus-informed course books, and skills books in an appendix. Meyer (2017) explores what makes corpus-informed materials different and why teachers and learners should use them. In her blog post, she explains what a corpus is and how it can be helpful in answering questions like “What are the most frequent words and phrases in English?” or “How often do people use idiomatic expressions and why?” She also explains how the
analysis of learner corpora can help teachers prioritize what to focus on in their teaching time. She not only stresses the fact that corpora lend themselves particularly well to the study of collocation (and gives concrete examples) but she also insists on the importance of developing learners’ awareness and familiarity with MWI from the early stages of language acquisition onwards (for more details, see www.cambridge.org/elt/blog/2017/04/24/corpus-informed-materials/).

Besides textbooks, some online resources provide learners and teachers with information on MWI. Some of these resources are simply lists of collocations, as is the case of the academic collocation list available in .pdf or .xls formats at https://pearsonpte.com/wp-content/uploads/2014/07/AcademicCollocationList.pdf, which contains over 2,400 academic collocations. For instance the 2,241st most frequent combination is (be) strongly correlated (with), 2,242nd is strongly disagree, and 2,243rd is (be) strongly influenced (by).

As indicated on the website (https://pearsonpte.com/organizations/researchers/academic-collocation-list/), the academic collocation list contains the 2,469 most frequent and pedagogically relevant lexical collocations in written academic English and was compiled from the written curricular component of the Pearson International Corpus of Academic English (PICAE) comprising over 25 million words. The development involved four stages: a computational analysis of the corpus, the refinement of the data-driven list based on quantitative and qualitative parameters, an expert review, and finally a systematization of the results presented. Expert judgment was used to ensure pedagogical relevance and usability of the results (cross-disciplinary collocations were for instance favored). As explained by the authors of the list, it can help learners increase their collocational competence but can also support EAP teachers in their lesson planning.

While the .xls format of the list allows for filtering and sorting, such lists remain rather static in format and may not be ideal in a learning-for-use perspective. Other resources offer a more flexible approach. One such free resource is SkELL, which stands for Sketch Engine for Language Learning (see Baisa and Suchomel (2014) for a lengthier discussion of the tool). As indicated on the SkELL website (see www.sketchengine.co.uk/skell/), it is a simple tool for students and teachers of English that can be used to easily check how a particular phrase or a word is used by real speakers of English. Behind the user-friendly interface in which the user simply keys in a search word (see http://skell.sketchengine.co.uk/run.cgi/skell), the tool searches a corpus of more than one billion words in English (news, scientific papers, Wikipedia articles, fiction books, web pages, and blogs). If the user clicks on Wordsketch, a list of words which occur frequently together with the searched word appears, and those words are sorted into various categories. A Wordsketch output of a search for the keyword food (with context button on) will provide users with categories such as “verb with food as object” (as in eat/cook/prepare/consume/grow food), with “adjectives with food” (as in food was delicious, food is plentiful, foods rich in . . . ), or with “nouns modified by food” (as in food supply, food allergies, food poisoning, food processor).

If the user wants more context, clicking on the desired word combination (e.g., food allergies) generates example sentences of the MWI selected found in corpora, as for instance:

- The role of food allergy is controversial.
- Food allergies seem far less common in underdeveloped countries.
- Food allergies are another major health concern with genetically engineered foods.

As the search engine uses fuzzy matching techniques and works with lemmatized, part-of-speech tagged, and partially parsed corpora, both singular and plural forms of food allergy
are provided. If the food + serve MWI had been selected, the user would have been provided with examples of active and passive sentences with the verb in potentially different forms, such as in the following examples also extracted from SkELL:

- An excellent restaurant serving traditional food and wine.
- The delicious food served was a daily feast.
- Food is served in small baskets.
- Food was beautifully served.

A tool like SkELL (with its rather intuitive interface) may be used by all types of learners while they are writing to help them find, for instance, frequent collocations (delicious, excellent, fantastic food). The word sketch tab gives a quick overview of frequent combinations. As for the examples tab, it offers up to 40 example sentences for each search, which allows slightly more advanced users to spot patterns of use. In the preceding examples, food + serve appears more often in passive constructions of the type food is served than in active constructions (someone serves food). Being able to spot patterns of use requires however, a slightly more complex competence needing at least some basic training (see the next section for a discussion of teacher-led DDL activities).

The SkELL interface is also available for languages other than English (at the time of writing, Russian and Czech). If (more advanced) users want to work on other languages, or use more complex search criteria they can use Sketch Engine, which is not freely available but which offers more options (see www.sketchengine.co.uk/), as it explores how language works by using complex algorithms to “analyse authentic texts of billions of words (text corpora) to identify instantly what is typical in language and what is rare, unusual or emerging usage. . . . [It] is used by linguists, lexicographers, translators, students and teachers . . . [and] contains 400 ready-to-use corpora in 90+ languages, each having a size of up to 20 billion words to provide a truly representative sample of language”.

Another example of an online tool that allows users to highlight MWI in a text is IDIOM Search (see http://idiomsearch.lsti.ucl.ac.be/). The online interface allows users to test a new algorithm for extracting the most set phrases from a text ranging from simple collocations and proper nouns (named entities) to idioms and proverbs (for the scientific background of the IDIOM Search project, see Colson, 2016). The beta version is available for Chinese, French, Spanish, and English, and the corpora used for the extraction of phrases are web corpora of about 200 million tokens. The noise in the output is an estimated 10%, varying from one language to another. Users have to select their language, copy paste the text to analyze, and the results page displays the set phrases identified in the input text. Colors correspond to the degree of fixedness and range from light shading (partly fixed) to dark shading (very fixed) combined with information on frequency (see Figure 22.1). Two ratios are also provided: a PW ratio corresponding to the number of phrases divided by the number of tokens, and a PT ratio which corresponds to the percentage of word tokens that are in the shaded zones. For example, a PT ratio of 0.42 indicates that 58% of the words are not part of a set phrase. Figure 22.1 presents the result of the query for the first three paragraphs of Alice in Wonderland (by Lewis Carroll in 1865).

While SkELL can potentially be used by learners as a reference tool to find collocations while they are writing a text for instance, IDIOM Search is more of a diagnostic tool as it somehow assesses the idiomaticity of a text. A preliminary study (Meunier & Bulon, forthcoming) shows that IDIOM Search can be useful is assessing the phraselogical competence of lower to upper intermediate learners of English and that the PT ratio provided can help
discriminate between levels of proficiency (0.66 correlation, with a rise up to 0.87 when a few other complexity measures are added). Teachers and learners could for instance use IDIOM Search to get a first impression of the types and proportion of MWI in their texts, and compare them to similar text types produced by other learners or by native speakers with a view of using the information presented in the comparison to adapt or edit their text.

The examples just listed are only some of the available tools that illustrate how (Learner) Corpus Research (L)CR has led to a new perception and representation of language in general, and of MWI in particular. Such examples account for the teaching content knowledge part of the Technological Pedagogical Content Knowledge model of MWI. One drawback, however, is that such tools may sometimes be used in a way that may lead to inert knowledge (as mentioned in the first section of the chapter) as the focus may be on “content” learning – e.g., memorization of MWI for a particular test without a focus on having learners retrieve those MWI for when they need them productively at a later stage.

In the next two sections I will comment on activities that foster process learning and attend to the second and third tenets of learning-for-use, which involves activities promoting goal-directed processes that put learners in situations where knowledge is constructed, used, and made accessible for future use.

### Technical and Pedagogical Knowledge: Data-Driven Learning

As a reminder, technical and pedagogical knowledge (TPK) refers to how teaching and learning can change when particular technologies are used in particular ways. Some technological affordances of (L)CR have made it possible to revisit the pedagogical approaches

---

**IDIOM Search Results**

Alice was beginning to get very tired of sitting by her sister on the river bank, and of having nothing to do. Once or twice she looked into the book her sister was reading, but it had no pictures or conversations in it. "What is the use of a book without pictures or conversations?", thought Alice. So she was considering the wise words she could, because the hot day made her feel very sleepy and stupid, whether the pleasure of making a daisy chain would be worth the trouble of getting up and picking the daisies. Suddenly a White Rabbit with pink eyes ran close by her. There was nothing so very remarkable in that, nor did Alice think it very much out of the way to hear the rabbit say to itself "Oh dear! Oh dear! I shall be too late!" (When she thought about it afterward, it occurred to her she ought to have wondered about this, but at that time it all seemed quite natural). But when the rabbit actually took a watch out of its pocket, and looked at it, Alice realised she had never before seen a rabbit with either pocket, or a watch to take out of it. She ran across the field after the rabbit, and had just in time to see it pop down a large rabbit-hole under the hedge.

**Figure 22.1** IDIOM Search results output

*Source: Illustration reproduced with permission of J.P. Colson*
Resources for Learning Multiword Items

used in foreign language learning. They include, among others, data-driven learning (DDL) approaches and tools. The term data-driven learning was coined in Johns (1991) and mainly refers to the fact that learners and/or teachers can work with/manipulate/query/use corpus data for language awareness-raising tasks (see Chambers, 2010, for a detailed history of data-driven learning, Lefko-Szymańska & Boulton, 2015, Boulton, 2017, and Boulton & Cobb, 2017, for in-depth discussions and illustrations of the potential of data-driven learning in language learning).

Data-driven learning activities are essentially a new form of consciousness-raising tasks exploiting the affordances of corpus tools. The corpus provides the data from which learners discover patterns for themselves, and the tools (concordancers mainly) act as some sort of combined input flood and input enhancement providers; see concrete illustrations and screenshots in the next section. A concordancer lists all of the examples of a keyword from a corpus or text.

Teachers can also manipulate the concordances to scaffold the awareness-raising activities (sorting the word on the right of the initial search word to make typical collocations more salient), see Mishan and Timmis (2015, pp. 81–82) for a concrete illustration of a scaffolded DDL activity. In such cases, teachers act as facilitators/guides in providing a scaffolded approach to the analysis of the data. Mishan and Timmins (ibid., p. 81) state that data-driven learning “places responsibility on learners to discover patterns in the data presented”. In addition, the authors note the similarities between data-driven learning and communicative language teaching as learners are seen as “active constructors” of knowledge instead of being passive recipients. The cognitive and analytical skills practiced and developed in data-driven learning activities offer more autonomy to learners and can (hopefully) become transferrable process skills. Readers interested in the impact of input flood on language proficiency are invited to read Mahvelati and Mukundan’s (2012) study of the role of participants’ cognitive style in collocational knowledge development. Another recommended article is Vyatkina’s (2016) study of the effects of data-driven learning of German lexico-grammatical constructions by North American college students with intermediate foreign language proficiency (see also Further Reading section at the end of the chapter). It should also be added here that data-driven learning activities, while generally useful, also have drawbacks (such as the loss of a larger context) and may not be beneficial to all types of learners. For a lengthier discussion of the limitations and potential drawbacks of data-driven learning, see Mishan and Timmis (2015) and Boulton (2017).

Illustrations: Some Online Tools Fostering Process Learning Activities

Data-driven learning activities are not necessarily computer-based and “paper-based” data-driven learning activities have also been shown to be effective for learners (see Vyatkina, 2016 for examples of both paper- and computer-based activities). As the present chapter focuses on the Technological Pedagogical Content Knowledge model however, I will only include digital open educational resources (OERs) in the present section (for more information on OERs, see www.unesco.org/new/en/communication-and-information/access-to-knowledge/open-educational-resources/what-are-open-educational-resources-oers/). Dudenev, Hockly, and Pegrum (2013) list the three main types of factors affecting digital activity choice: pedagogical (students as language learners), personal (students as individuals), and digital (students as technology users). Pedagogical factors include, among other things, class context (native, second, or foreign language environment), language level (linguistic proficiency of the students) and language needs (particular language needs of the
students). Personal factors are related to age, interests, and cultures of the students; and digital factors refer to students’ attitudes towards digital technologies, their technological proficiency and confidence, their digital literacy levels, and the equipment and tools students (and teachers) have access to.

Masterman and Wild (2011), in turn, focus on the teacher’s perspective and list factors that are critical in whether or not teachers use open educational resources. They include the relevance of content and fit to the lecturer’s current purpose, and provenance and pedagogic intent as factors. Masterman and Wild also analyze the logistical factors that can account for teachers’ willingness to work with OER, and these include among others: technical and implementation issues; a conceptualization of teaching as helping students to become active, independent learners; and readiness to learn (i.e., continuous professional development practice).

As can be seen from the two preceding paragraphs, a sustainable use of digital tools by both learners and teachers is far from being undemanding, especially if all the learner- and teacher-related factors have to be met simultaneously. The two tools that will be presented next have the potential to address learner- and teacher-related factors but the actual use that will be made of the tools – as well as their potential long-term benefits for learners – cannot be assessed or predicted once and for all. They will be very much related to specific contextual elements.

The first tool is the Word and Phrase Info tool, available at www.wordandphrase.info/analyzeText.asp. This tool allows users to enter any text and see useful information about words and phrases in the text. The information accessed is based on data from the Corpus of Contemporary American English (COCA). In addition to information at the single-word level (e.g., highlighting of all of the medium and lower-frequency words in the text, or of academic words) clicking on any word in the text gives access to a detailed word sketch with re-sortable concordance lines, and frequency of the word (overall, and by genre). The Word and Phrase Info tool allows for searches on selected phrases in the text (and provides related phrases in COCA). This resource is defined as a “collocational thesaurus” by its creator (Mark Davies). Figure 22.2a shows a screenshot of a sample of text analyzed by Word and Phrase. In the example selected, I clicked on the word *embedded* in the text and was provided by a list of concordances of the word in several text types.

As can be seen, *embed* is particularly frequent in academic English (when compared to spoken English for instance). In addition, a quick look at the concordances lines (only seven are visible in Figure 22.2a, but a total of 201 concordance lines were produced) shows that the preferred structure is a passive one, followed by *in* as a preposition (121 instances out of the 201 displayed) – see Figure 22.2b for a screenshot of the concordances and the ease of spotting frequent structures thanks to the colors and the pre-sorted concordance lines.

As for phrases, Davies also explains that clicking on the groups of words (e.g., *potent argument* in the text entered), will provide the user with alternate ways of expression (e.g., *powerful* or *convincing* argument), and will show the frequency of those phrases in COCA. This helps users find just the right phrase based on a huge collection of native English texts (see www.wordandphrase.info/analyzeText.asp).

The second tool is the Louvain English for Academic Purposes Dictionary (LEAD). Paquot (2012, p. 165) defines LEAD as “an integrated dictionary and corpus tool intended to help non-native speakers write academic texts in English”. Corpus data in LEAD is meant to be customized “to users’ needs in terms of discipline and mother tongue background”. LEAD is both a dictionary and a writing-aid tool that when used successfully should contribute to
Resources for Learning Multiword Items

Figures 22.2a and 22.2b

Learning. LEAD offers options to search for English academic words and expressions (see Paquot, 2012, for illustrations and screenshots of the tool). It can be used to access:

- Definitions and examples of word use in context (preferred position in the sentence, typical collocations and recurrent phrases),
- Charts to help users remember salient features (e.g., frequency differences across genres),
- Warnings against frequent errors, and usage notes that compare learner and expert writing,
- Context-sensitive lexicographical treatment of phraseology, as corpora of expert and learner writing have been used in the compilation of the tool,
- Numerous exercises on various EAP language functions.

Interestingly, Granger and Paquot (2015) have recently conducted a needs analysis of users of LEAD. The results indicate that a large proportion of the users (80%) believe that they will use this tool for future assignment, thereby pointing towards one of the key tenets of learning-for-use: accessibility for future use. Students liked the hyperlinked discipline-specific...
examples. However, the direct access to a concordancer was not particularly well-rated by MA students, which contrasted with PhD students’ appreciation of the concordancing p-facility.

Teachers’ perceptions of online tools is also an important predictor of their future inclusion in the classroom. The assessment of a project aiming to foster teachers’ uptake of OERs including natural language processing aspects (the TELL-OP project, see www.tellop.eu/ for more details) has revealed the importance of pre- and in-service teacher training in the use of such digital tools. The 169 teachers (spread in four cohorts over four countries: Belgium, Germany, Spain, and the UK) who followed the online module set up for the teacher training output of the project showed a lot of interest in the teaching opportunities provided by the different natural language processing tools presented. They explained that such tools would indeed enable learners to gain autonomy in their learning process (Meunier, 2017). Despite the interest expressed, however, some of the corpus tools and data-driven learning activities proposed generated tepid reactions on average, even though they scored slightly better among mobile-trained respondents and teachers working in a mobile-friendly environment (Meurice, 2018; Pérez-Paredes, Ordoñana Guillamón, & Aguado Jiménez, 2018). As is (still too) often the case in studies presenting tools and approaches for data-driven learning, the main focus is on training teachers in how to use specific tools. However, there is still not enough focus on what Koehler and Mishra (2005) call a learning by design approach, something that will be further discussed in the next section.

**Future Directions**

The second section of the chapter has, I hope, shown that access to MWI in situated use is now possible thanks to (L)CR and its multiple features. It has also been shown that some tools are more static and some are more dynamic; that some tools are better suited for assessment purposes (how phraseological is my text) and some others are more useful for productive tasks (what are the possible collocations for the specific word I want to use). It is also obvious from the various examples presented that not all tools can retrieve all types of MWI, and that each tool often uses its own terminology. For example, IDIOM Search (see Figure 22.1) uses very broad categories, whereas a tool like LEAD will provide users with more complex search options including specific parts-of-speech. While some users may regret the lack of terminological homogeneity, I do not necessarily find it problematic from a learning-for-use perspective. I would even go as far as saying that too strong a focus on meta-language (is this a communicative phraseme? is this a fixed or semi-fixed idiom? is this a phrasal verb or a prepositional verb?) obscures the value of tools and makes both teachers and learners shy away from data-driven learning that they find either too specific or too linguistically complex. This is particularly the case for instructors teaching learners with lower levels of proficiency. I see the overall added-value of each of the tools presented here as serving the cause of revealing the highly conventional nature of language to both teachers and learners and as key resources in understanding and accessing frequency information or form-meaning mappings in use. I also hope that researchers in (L)CR and natural language processing circles will go developing and refining such tools in the future.

With time and experience in pre- and in-service teacher training, I have also come to realize that advertising tools is just as important as explaining what they could actually do. While in the past I tended to present “corpus linguistics tools” to teachers, I now present them as “get your 24/7 native speaker assistant for free” (Meunier, Vincent, Suner, & Van de Vyver, 2016). Such a small verbal move has made a huge difference in how teachers approach the
tools and their features, and I can recommend using similar analogies to help potential users to see the benefits of data-driven learning in pre- and in-service teacher training circles.

The importance of learner- and teacher-oriented pedagogical, personal and digital factors (Dudeney et al., 2013, and Masterman & Wild, 2011) that are critical in whether or not teachers use open educational resources should not be underestimated. The actual use that will be made of the exiting tools, their future developments, and their potential long-term benefits for learners will depend on the awareness of, attention to, and integration of such factors. As the Technological Pedagogical Content Knowledge model is the “understanding that emerges from interactions among content, pedagogy, and technology knowledge” (Koehler et al., 2013, p. 16), I second Koehler and Mishra’s (2005) plea for a learning by design approach to the teaching and learning of MWI. This may be achieved by helping both teachers and learners develop “an understanding of the complex set of interrelationships between artifacts, users, tools, and practices” as it will promote a “flexible and situated understanding of technology” (Koehler & Mishra, 2005, p. 94).

I would like to finish this chapter by prompting teachers and learners alike to use existing Open Educational Resources to get to grips with MWI. One such example is a recently created SPOC (Small Private Online Course) from the University of Queensland (Australia) entitled “Improving Writing Through Corpora: Data-Driven Learning”. This short (five-hour) course, created by Peter Crosthwaite (see https://edge.edx.org/courses/course-v1:UQx+SLATx+2018_S2/about), aims to provide its users with the tools, knowledge, and skills to improve their writing. While the course contains functions that are only available to UQ students (marking of tasks for instance), it is open to anyone who would like to know more about corpora and some freely available collocation tools. Despite the fact that each course will probably have its strengths and weaknesses and target more or less specific audiences, sharing good practices through open educational resources seems to be the right way to help promote Technological Pedagogical Content Knowledge.

Further Reading


This study uses systematic meta-analytic procedures to summarize findings from experimental and quasi-experimental investigations into the effectiveness of using the tools and techniques of corpus linguistics for second language learning or use. Sixty-four separate studies representing 88 unique samples are covered. Large overall effects for both control/experimental group comparisons and for pretest/posttest designs are reported. The authors also explain that while DDL research has improved over the period investigated, changes in practice and reporting are recommended.


This volume explores how corpus data can be used as part of the learning process. It provides research insights based in the classroom context and reports on several state-of-the-art projects around the world. It also addresses issues involving different types of corpora, for different learner profiles, and for different purposes.


This brand new book offers a multifocal perspective on formulaic language from a second language acquisition perspective. It addresses cognitive, psycholinguistic, sociocultural, pragmatic, and pedagogical aspects. The six chapters of Part 3 of the book will be particularly relevant for readers with pedagogical interests.

This paper explores the effects of data-driven learning of German lexico-grammatical constructions and compares the effects of computer-based and paper-based activities in learners’ immediate and delayed performance gains. It also explores changes in learners’ proficiency and perceptions of data-driven learning. It is one of the rare studies focusing on a second language other than English, and it also combines the different outcome measures in a multilevel modeling design.

Related Topics
Classifying and identifying formulaic language, factors affecting the learning of multiword items, resources for learning single-word items, key issues in teaching multiword items

References


