Lang. Teach. (2020), **53.1**, 29–43
doi:10.1017/S0261444818000277 First published online 25 September 2018

Plenary Speech

Combining user needs, lexicographic data and digital writing environments

Ana Frankenberg-Garcia University of Surrey, UK
a.frankenberg-garcia@surrey.ac.uk

The past decades have seen dramatic improvements to dictionary content and format. Yet dictionaries – both paper-based and digital – remain disappointingly underused. As a result, it is widely acknowledged that more needs to be done to train people in dictionary-consultation skills. Another solution would be to build lexicographic resources that require little or no instruction. In this paper, I present the ColloCaid project, whose aim is to develop a lexicographic tool that combines user needs, lexicographic data and digital writing environments to bring dictionaries to writers instead of waiting for them to get the information they need from dictionaries. Our focus is on helping writers produce more idiomatic texts by integrating lexicographic data on collocations into text editors in a way that does not distract them from their writing. A distinguishing characteristic of ColloCaid is that it is not limited to providing feedback on miscollocations. It also aims to 'feed forward', raising awareness of collocations writers may not remember or know how to look up. While our initial prototype is being developed specifically for academic English, the implications of our research can be broadened to other languages and usages beyond academic.

1. Introduction

When I first started teaching, my students used to bring trolleys-full of dictionaries on exam days. Today's students look up words on their computers or mobile phones instead. It is not just the medium that has changed, however. In terms of content, current state-of-the-art learner dictionaries contain information that goes well beyond spellings, definitions or translations, providing users with valuable empirical, corpus-based data on how to employ words in texts. This includes information on word grammar, lexical collocations and common errors to avoid, as well as typical examples of usage.

Yet despite the remarkable developments that have taken place in the field of lexicography over the past decades, dictionary-user behaviour does not seem to have changed much. People tend to turn to dictionaries mostly to look up meanings, spellings, translations, or as an authority when engaging in games such as crosswords (Atkins & Varantola 1997;

Revised version of a plenary address given at TechLing 2017, University of Bologna, Forlì, Italy, 10–11 November 2017.

Frankenberg-Garcia 2005; Welker 2006; Frankenberg-Garcia 2011; Lew & de Schryver 2014; Müller-Spitzer 2014; Nesi 2014; Gromann & Schnitzer 2016; Jardim 2018). Little do average dictionary users know that they could also consult dictionaries to help them use language more idiomatically.

Considering this reality, it is widely acknowledged that more needs to be done to teach dictionary-consultation skills. However, in an age where authoritative dictionaries are rapidly losing ground to easily accessible free, online language tools and resources (Levy & Steel 2015; Lew 2016), it would be hard to convince the public in general of the advantages of learning to become better users of dictionaries.

In this paper, I propose bringing lexicographic information to writers instead of waiting for them to get the information they need from dictionaries. I begin with an overview of how pedagogical dictionaries have evolved, and of some of the challenges of getting writers to become better users of dictionaries and of assisting writers in real-time. Next, I present the ColloCaid project, which combines user needs, lexicographic data and digital writing environments to help writers produce more idiomatic texts.

2. Developments in pedagogical lexicography

Dictionaries in the past were mostly a repository of the words in a language, with a focus on definitions, the standardization of spellings, and etymology (Cowie 2009). In addition, bilingual dictionaries provided translations, but neither monolingual nor bilingual dictionaries were particularly concerned with usability. It was up to the users consulting these resources to decipher their 'cryptic lexicographic content' (Lew & de Schryver 2014: 341). Interest in the pedagogical function of dictionaries, where the end user began to be taken into account, saw the development of new types of dictionaries for learners. The *Idiomatic and syntactic dictionary of English* (Hornby, Gatenby & Wakefield 1942) – the precursor to the famous 1974 edition of the *Oxford advanced learner's dictionary* (OALD) (Hornby, Cowie & Lewis 1974) – is regarded as the first dictionary to address information such as noun countability and verb complementation, which can help learners use words in language production tasks. The hugely popular 1974 edition of the OALD then added phonetic transcriptions to aid pronunciation and examples to illustrate usage (Cowie 1999). Another significant development in the field of pedagogical lexicography was the introduction of a controlled defining vocabulary in the 1978 edition of the *Longman dictionary of contemporary English* (Procter 1978), where a conscious effort was made to restrict the words used in definitions to those learners are more familiar with, thus increasing the chances of users understanding the meanings of the words they consult without having to look any further.

The next paradigm shift in the history of dictionaries for learners occurred with the publication of the *Collins COBUILD English dictionary for advanced learners* (Sinclair 1987a). Among other innovations, such as providing more natural-sounding, full-sentence definitions, COBUILD was the first dictionary to be compiled with the support of a computerized corpus with millions of words of English used in authentic communicative situations. Whereas up to then lexicographers had had to rely on their own partial perceptions and experience, the corpus

https://doi.org/10.1017/S0261444818000277 Published online by Cambridge University Press

express, protect, show, pique, spark, represent, lose, serve, pay, pursue, generate, charge, share, attract, accrue, grow, have, register, earn, advance, increase, promote, arouse, calculate, compete, defend, peak, stimulate, reflect, develop, indicate, safeguard, further, capture, demonstrate, hold, take, draw, catch, add, maintain, gain, compound, deduct, suit, acquire, align, renew, retain, balance

Figure 1 Top 50 verbal collocates of the noun *interest* in the enTenTen13 corpus

revolution took language description to new levels, enabling them to capture the combined intuitions of hundreds or even thousands of language users together. Corpus software counts, sorts, ranks and displays words in special ways that facilitate linguistic analysis. For the first time, lexicographers could describe a representative selection of empirical language data systematically. With corpora, it also became possible to take word frequencies into account when establishing defining vocabularies and deciding which senses were more important to present to learners (Sinclair 1987b). Since the words in a language tend to follow a Zipfian distribution (Zipf 1949), where the top-ranking words cover most of the language (Nation 2001), it made sense for dictionaries for learners to prioritize more frequent words and senses.

Corpora also allow lexicographers to analyse how words are used together. This enables them to provide learners with information on not only syntactic patterns or grammatical collocations (e.g. *interest* IN *something*), but also on lexical collocations, i.e. conventional combinations of lexis like *EXPRESS/TAKE/SHOW an interest in something*, which make texts sound natural and idiomatic (Nattinger & DeCarrico 1992; Hoey 2005; Nesselhauf 2005; Paquot & Granger 2012; Wray 2013; Boers & Webb 2017).

If you ask experienced language users about lexical collocations, say, what verbs can be used before the noun *interest*, they may recall around two or three without hesitation, but usually need to think harder to remember more (Frankenberg-Garcia 2018). With corpora, however, it is possible to extract a long list of verbs that collocate with *interest* as an object in seconds, as exemplified in Figure 1 with data from the 20 billion-word enTenTen13 corpus, available on Sketch Engine (Kilgarriff et al. 2014). Without getting into too many details, this is done by comparing the overall frequency with which words appear in a corpus (e.g. *express/show/take*) with the frequency with which they appear in proximity to a target word (in this case, *interest*), and calculating the likelihood of the two appearing together.

This not only greatly facilitates the work of lexicographers, but is also especially relevant to learners, for collocations have been shown to be particularly difficult to master (Nattinger & DeCarrico 1992; Nesselhauf 2005; Paquot & Granger 2012; Wray 2013; Boers & Webb 2017). Since texts that do not make use of appropriate collocations tend to sound less fluent/proficient (Hsu 2007; Crossley, Salsbury & McNamara 2015) and are notoriously more difficult to process (Hoey 2005; Ellis, Simpson-Vlach & Maynard 2008; Conklin & Schmitt 2012), the inclusion of information on collocation in dictionaries represents a particularly welcome innovation to help learners use language more idiomatically.

Another significant change brought about by corpora was the replacement of scattered examples to illustrate definitions with a more consistent use of authentic, corpus-based examples selected to further clarify meaning or draw attention to typical usages of words in context (Sinclair 1987b). This development was extremely important, as

https://doi.org/10.1017/S0261444818000277 Published online by Cambridge University Press

dictionary-use research has shown that examples help learners with language production (Frankenberg-Garcia 2012a, 2014, 2015).

Following the corpus revolution, the next leap in pedagogical lexicography took place with the popularization of personal computers, and the possibilities offered by presenting dictionary information in a new medium (Lew & de Schryver 2014). Rather than simply transposing the print editions of dictionaries to digital formats – initially as CD-ROMs – the major English dictionaries for learners introduced several innovations. To begin with, finding words has become much easier. Users are no longer required to look up words in alphabetical order, as they can now just type them into a search box. If users do not know exact spellings, they only need to begin typing to be reminded of matching words, or corrected spellings when a word is misspelled. Similarly, users do not have to know the uninflected forms of words to look them up, as inflections are recognized too. The new medium has also made it easier for users to learn how to pronounce words, as they can click on sound files to listen to them instead of having to decipher phonetic symbols. Recently, sound files have also begun to be used to enhance the definitions of words involving sounds. For example, in the *Macmillan English dictionary online* (Rundell 2009) entry for the verb *bark*, one can click on a sound icon to hear a dog barking.

Another particularly important advantage of the new electronic medium is space (Lew & de Schryver 2014; Rundell 2015). The fact that dictionaries do not need to be printed anymore has meant they are no longer restricted by the weight and cost of paper or of using colour. There is therefore room for unpacking the compact way in which information is traditionally presented in print editions, enhancing the retention of information (Dziemianko 2015, 2017; Choi 2017). There is also room for expanding contents, like adding further examples of usage, as well as vocabulary exercises and games (Lew 2011). However, it could be argued that space is only an advantage if used wisely. Overburdening dictionary users with too much information could be detrimental, as it could make look-ups less efficient or even distract users from the reason why they were consulting a dictionary in the first place.

More recently, with the proliferation of wireless internet access and portable electronic devices, there has been a growing tendency for electronic dictionaries to migrate from static CD-ROM versions to online platforms which can be accessed remotely. In addition to the obvious benefits of portability, the move to online dictionaries has paved the way for further developments in the field. First, dictionaries 'can be updated as often as needed, and all users can instantly benefit from the improved content or features right from the moment these become available' (Lew & de Schryver 2014: 345). Another advantage of online dictionaries is that log files can offer new insights into user behaviour, which can in turn be fed back into the development of subsequent dictionary updates (de Schryver & Joffe 2004). However, this is more easily said than done, since log files tell us little about the motivations behind individual look-ups or the users themselves (Santos & Frankenberg-Garcia 2007). Moreover, in addition to privacy issues, the major players in pedagogical lexicography seem to have kept this type of data to themselves, as there does not seem to be much published research on dictionary-user log files and how they can promote better dictionaries.[1] On the other hand, this does not mean to say there is no attempt to gain information from actual users. In

[1] However, see Müller-Spitzer, Wolfer & Koplenig (2015) for a recent study on log files from the German version of Wiktionary.

https://doi.org/10.1017/S0261444818000277 Published online by Cambridge University Press

fact, it is now common for online dictionaries to encourage user-generated content (Rundell 2017). A notable initiative is the *Macmillan English dictionary online* (Rundell 2009), where users are invited to contribute to the addition of new entries whenever they look up words that are yet not part of the dictionary's headword list.

3. Getting writers to use dictionaries

Despite the spectacular advances in dictionary content and format outlined in the previous section, as pointed out in the introduction, dictionaries remain by and large underused, particularly as an aid to writing. I have just come back from examining a Ph.D. thesis on dictionary use (Jardim 2018), and its findings confirm yet again previous research showing that users are generally unaware that dictionaries are not just about meanings, spellings, settling language disputes or L1-L2 equivalence (see Introduction). Few writers realize that dictionaries can also help them use words in context and produce more idiomatic texts.

Although existing research recognizes the need to train users in dictionary-consultation skills (Frankenberg-Garcia 2011; Ranalli 2013; Kim 2017), the aforementioned studies on dictionary use show that little progress has been made in this arena. What is particularly worrying is the inadequate way in which information about dictionaries is being conveyed to the general public even today. For example, the top result for a quick online search for 'how to use a dictionary' carried out when preparing this paper took me to wikihow.com, which outlined the following steps:

a. *Choose the right dictionary*
b. *Read the introduction*
c. *Learn the abbreviations*
d. *Learn the guide to pronunciation*
e. *Find the section of your dictionary with the first letter of your word*
f. *Read the guide words [i.e. the running head showing the first and last word on each page]*
g. *Scan down the page for your word*
h. *Read the definition*
i. *Alternately, you could use an online dictionary*

As can be seen, apart from point (i), the above instructions are totally out of step with recent developments in the field. Yet even if dictionary users were made aware that dictionaries have evolved not just in terms of format, the fact is most language users are not in the habit of consulting references to help them become better writers. As explained in Frankenberg-Garcia (2014: 140), 'one of the main reasons why learners are underusers of dictionaries and other language resources is that they are often not aware of their own language limitations and reference needs'. While people normally realize when language comprehension is an issue, they tend to be less aware about language production problems. In a second language writing workshop at a Portuguese university, where undergraduate students were encouraged to ask for help at any point during writing, 'the queries posed by the students suggested that they felt all they needed to become successful writers of English

https://doi.org/10.1017/S0261444818000277 Published online by Cambridge University Press

was a bilingual dictionary and a spelling checker' (Frankenberg-Garcia 1999: 104), although the problems in the texts they produced went far beyond that.

Promoting better dictionary-consultation skills among writers cannot have much impact if they are not sufficiently aware of their reference needs in the first place. Moreover, in an age where authoritative dictionaries are competing with other types of language tools and resources (Levy & Steel 2015; Lew 2016), the time is ripe for rethinking pedagogical lexicography. In the next section, I propose bringing lexicographic information to writers instead of expecting them to get the information they need from dictionaries.

4. Bringing dictionaries to writers

While writers know they can look up translations and spellings in dictionaries or dictionary-like tools, one of the greatest challenges of pedagogical lexicography is to get them to use dictionaries for more than that. Collocations are particularly relevant in the context of writing. As referred to in Section 2, collocations have been shown to be notoriously difficult for language learners. Failing to follow the established collocation conventions of a particular language or language variety can lead to error (e.g. **based IN something*, **to LEARN knowledge*) or less idiomatic text that can be harder to process. For example, compare the collocation *DEEPLY entrenched*, which proficient language users tend to read as a unit, with a less idiomatic combination of words like *INCREDIBLY entrenched*.

There are many references writers can consult when in doubt about collocations in English. In addition to looking them up in corpus-based, general learner dictionaries, there are also dictionaries that focus specifically on collocations, like the *BBI dictionary of English word combinations* (Benson, Benson & Ilson 1986), the *Oxford collocations dictionary* (Runcie 2002), the *Macmillan collocations dictionary* (Rundell 2010), and the *Longman collocations dictionary and thesaurus* (Mayor 2013).

Language users can also look up English collocations in free online tools like *Just the Word* and the *Flax Library*, which process corpus data and provide automatic summaries of collocations. In addition, although there are not many language users familiar with corpora (Frankenberg-Garcia 2012b), those who are can go directly to the sources where lexicographers get information about collocations in the first place. The British National Corpus (BNC), the Corpus of Contemporary American English (COCA) and more recently Sketch Engine for English Language Learning (SkELL), for example, are all easily accessible corpora of General English which writers can consult to help them with their use of collocations.

While there is no room here for a comprehensive review of all existing collocation aids available for English, there is certainly no lack of resources that writers can utilize to look up ways to combine words so as to improve the idiomaticity of their texts. However, unlike using dictionaries to look up more obvious reference needs like how to say a word in another language or check its spelling, language learners would have no reason to look up collocations if they were not aware of their shortcomings regarding them. In a controlled experiment with Hebrew learners of English, Laufer (2011) found that they had a tendency to misjudge what

https://doi.org/10.1017/S0261444818000277 Published online by Cambridge University Press

they knew about collocations and did not think it necessary to consult dictionaries to look them up. Similar evidence was found in Frankenberg-Garcia (1999, 2014). Moreover, even if writers realized collocations were a problem, they would have to interrupt their writing to look them up and could lose their focus in the process, forgetting what they wanted to say. This can be particularly detrimental to writers struggling with cognitively demanding texts. In a study observing how academic writers interacted with online dictionaries and corpora, Yoon (2016: 220–221) observed that the participants 'expressed frustration with the time required to go through the consultation cycle' and complained that they 'had their flow of thoughts interrupted'.

A solution to this problem would be to provide writers with real-time help. Writing tools are becoming better and better, with various innovations that can assist writers on the spot. For example, most text editors today can autocorrect spelling or flag up misspelled words. Some writing tools allow users to right-click on a word to look up synonyms. Other useful functionalities include drawing attention to repeated words and missing punctuation. Researchers working on the Danish version of MS-Word are trialling an add-in that integrates a Danish-English bilingual dictionary and predictive text (Tarp, Fisker & Sepstrup 2017). In addition, recent developments in natural language processing and machine learning have given rise to sophisticated writing assistants like Grammarly®, which provide feedback on more complex issues such as verb tenses and word choice. Cambridge English has developed Write&Improve, which sets topics for non-native speakers of English to practise writing and gives them automatic feedback on their texts based on how similar texts were previously marked. Another well-known tool is Hemingway, which aims to inform writers on the readability of their texts based on sentence length and the use of adverbs and the passive voice or rarer words. WriteAway, in turn, processes data from corpora to autocomplete writers' sentences.

The provision of automated feedback on writing is a very fertile and fast-developing field, and it is hard to keep up with all the writing assistants that are emerging. While there are many truly innovative ways of helping writers in real time being proposed, the programmed advice given by some tools can at times be prescriptive and overly simplistic (e.g. 'avoid the passive voice'), and there does not seem to be enough research assessing the usability of these tools. Anyone who has used predictive text, for example, will know how annoying it can be. If predictive text can irritate users writing simple text messages on their phones, imagine its effect when writers are trying to cope with more cognitively demanding tasks. Another limitation is that if we exclude the integration of predictive text, writing assistants are mostly limited to offering corrective feedback. The challenge is thus to develop a lexicographic tool that is not just reactive, but which can also help writers proactively, without disrupting their writing. In the next section, I describe the ColloCaid project, which aims to bring collocations to writers instead of waiting for them to look up collocations they may not even be aware they need.

5. The ColloCaid project

ColloCaid is a three-year project led by myself at the University of Surrey, in collaboration with Professor Jonathan Roberts (Bangor University) and Professor Robert Lew (Adam

https://doi.org/10.1017/S0261444818000277 Published online by Cambridge University Press

Mickiewicz University), with the assistance of Dr Geraint Rees (Surrey University) and Dr Nirwan Sharma (Bangor University). The project is funded by the UK Arts and Humanities Research Council.

The principles underpinning our research apply to collocation in general, and in future we would like to see similar applications for various languages. However, given the limitations of what can be realistically achieved within the scope of three years, our prototype is being developed specifically to help novice users of English for Academic Purposes (EAP). This enables us to focus on the collocation needs of a well-defined group of real-world users for whom writing is particularly important.

5.1 User needs

The first step in our research was to identify the collocation needs of our target users. Based on the premise that there are no native speakers of academic language (Kosem 2010; Hyland & Shaw 2016; Frankenberg-Garcia 2018), ColloCaid aims to encourage novice EAP users (including native English speakers) to employ collocations which may not be instinctive to them. We nevertheless acknowledge that EAP users of different first language backgrounds may experience diverse problems regarding the use of collocations. It is well-documented that the ways in which second language writers combine words can be negatively impacted by their first languages (Nesselhauf 2005; Laufer & Waldman 2011; Peters 2016; Paquot 2017). By the same token, as shown in Frankenberg-Garcia (2018), less experienced native English EAP users tend to employ general language words and collocations which may sound out of place in formal academic writing. At a later stage in our research, we will therefore use learner corpora to analyse how such problems manifest themselves and provide targeted feedback to help writers tackle them.

However, before focusing on the comparatively more straightforward problem of corrective feedback, we intend to 'feed forward' first, addressing issues that are not as visible in learner corpora. The problem of collocations, after all, is not limited to error, but involves also the underuse and overuse of certain word combinations (Durrant & Schmitt 2009; Paquot 2017). At the root of such problems is not only the previously discussed tendency to overestimate knowledge of collocations (Section 4), but also lexical avoidance strategies (Faerch & Kasper 1983), whereby writers alter, reduce or completely abandon what they meant to say when they are unable to find the words they need. The starting point for the lexicographic database behind ColloCaid was therefore the identification of a core set of collocations that will be useful to EAP users, even if they themselves are not initially aware of their worth.

For this purpose, we opted to concentrate our efforts on collocations used across a range of academic disciplines. Without diminishing the importance of discipline-specific collocations, as discussed in Frankenberg-Garcia et al. (2018), we believe it is easier for EAP users to acquire such vocabulary incidentally, through 'a targeted and concentrated exposure to the subject-matter of their studies'. On the other hand, interdisciplinary academic collocations can be harder for novice EAP users to recall, precisely because they tend to be less salient. Although we do not rule out the development, at a later stage, of discipline-specific versions

https://doi.org/10.1017/S0261444818000277 Published online by Cambridge University Press

of ColloCaid, a writing assistant that handles interdisciplinary EAP collocations can be more immediately useful to a greater number of users.

5.2 Lexicographic data

Following the Zipfian principles referred to in Section 2, a combination of three well-established general EAP vocabulary lists was used to ensure appropriate coverage was given to words with the potential to improve the collocation repertoire of EAP users: the Academic Keyword List (Paquot 2010), the Academic Collocations List (Ackermann & Chen 2013) and a subset of the Academic Vocabulary List (Gardner & Davies 2014) identified by Durrant (2016) as being particularly relevant to novice writers. Since the three lists are based on different corpora and different extraction methods, combining them allows us to prioritize what they have in common. As detailed in Frankenberg-Garcia et al. (2018), our guiding principles in this selection were to focus on lemmas used across a wide range of disciplines (all three lists), including academic lemmas like *table* and *figure* that are also used in non-academic contexts (Academic Keyword List), prioritizing lemmas which evoke strong collocations (Academic Collocations List), and lemmas novice EAP writers actually use (Durrant's subset of the Academic Vocabulary List). The circa 500 noun, verb and adjective lemmas that overlapped in at least two of the three lists helped to determine which collocation nodes to focus on in the compilation of ColloCaid's initial lexicographic framework.[2]

The next step was to research lexical and grammatical collocates for the collocation nodes selected in corpora of expert academic writing. As explained in Frankenberg-Garcia et al. (2018), our main source was the 70 million-word Oxford Corpus of Academic English, which was kindly made available to our team on Sketch Engine, although we also consulted the Pearson International Corpus of Academic English (Ackermann, de Jong & Tugwell 2011) by kind permission of Pearson Longman and the academic components of the BNC and COCA. The analysis was centred on the logical collocation queries prompted by each node. For example, writers may ask questions like, 'What adjective can I use with *number*?', but are unlikely to ask, 'What noun can I use with *significant*?' because nouns are the foundation for the selection of adjectives, and not the other way around.

When deciding on which collocates to present, we opted to focus on collocations used across a range of academic disciplines. Therefore, discipline-specific collocations like *NATURAL/PRIME number* were left out, allowing us to devote more room to interdisciplinary academic collocations like *LARGE/INCREASING/SIGNIFICANT/AVERAGE number*. Following studies like Frankenberg-Garcia (2012a, 2014, 2015) on the value of examples for language production, our collocation framework is also being populated with corpus-based examples. These are being curated to: (a) show how collocations are used in context; (b) expose writers to further collocations (e.g. *LARGE number* followed by *OF*); (c) emphasize typical colligational, i.e., grammar, patterns (e.g. *INCREASING number, INCREASED numbers*); and (d) help users differentiate between semantically similar collocations like *INCREASING/GROWING number*.

[2] Adverb lemmas were disregarded as they do not normally prompt collocation queries. For example, writers would not normally ask themselves 'what adjectives can I use with *highly*?'.

https://doi.org/10.1017/S0261444818000277 Published online by Cambridge University Press

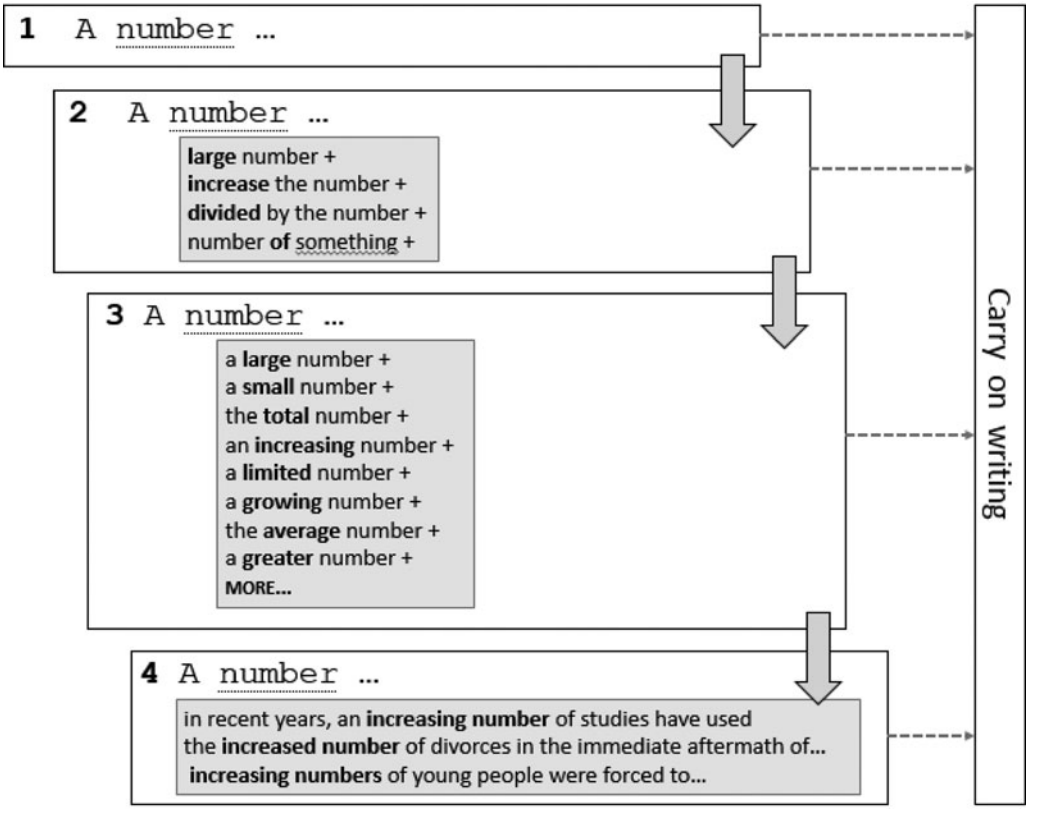

Figure 2 Incremental display of information on collocations

5.3 Digital writing environments

As discussed earlier, our main concern when integrating lexicographic data with digital writing environments in ColloCaid is not to distract writers from their writing, while helping them (a) not to give up on collocations through avoidance strategies; (b) find suitable collocates for the words they use without having to consult external resources; (c) notice collocations they may not remember to look up (because they overestimate their knowledge of collocations); and (d) self-correct miscollocations.

To facilitate the smooth integration of the interdisciplinary EAP collocation framework we are compiling, we want to flag up that relevant information on collocation is available in an unobtrusive way. This will be achieved by discreetly highlighting the lemmas that form part of our lexicographic framework in real time. Writers can then click to obtain further information or ignore and carry on writing. While our prototype is still under development, a schematic representation of how we propose to do this is shown in Figure 2. Step 1 in the figure illustrates how the highlighting of collocation nodes could be accomplished.

Should writers click on the highlighted lemma, they will be shown different academic collocations associated with it (Figure 2, Step 2). Note that instead of using metalanguage like

https://doi.org/10.1017/S0261444818000277 Published online by Cambridge University Press

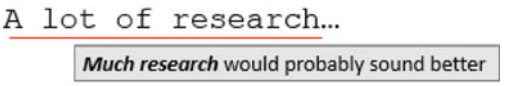

Figure 3 (Colour online) Suggesting more appropriate collocations

ADJECTIVE + *number*, *VERB* + *number*, and so on, we have chosen to present this information by displaying the strongest collocate pertaining to each grammatical relation.[3] This has the double advantage of sheltering less linguistically aware users from grammar and enabling writers to find the collocate they need without any further interaction.

If Step 2 is not enough, users can click on the plus sign to expand a grammatical relation with further options. Step 3 of Figure 2 shows the expansion of *LARGE number* with further adjectival collocates. Although in theory it would be possible to present a much greater number of suggestions at this point, we have opted to show a maximum of eight because of the known limitations of the working memory (Miller 1956), and also so as not to overcrowd the text-editor screen. The collocates displayed are the first eight in terms of logDice strength of association score.[4] However, they can click on *more* to view further collocates on a side-bar.

If writers still need more details, in the next interaction they can click on the plus sign to view corpus-based examples showing the selected collocation in context. Step 4 in Figure 2 illustrates this with the expansion of *an INCREASING number*. Note that unlike dictionaries, which normally give only one (if any) example to illustrate a particular collocation, we have opted to present three analogous examples, following research showing that multiple examples tend to help more (Frankenberg-Garcia 2012a, 2014, 2015). Since the examples have been curated to display further collocations and typical colligational patterns (see Section 5.2), it is likely that users will find one that can be transferred to their own texts with minimal adaptation. Note also the collocations in the examples are typographically enhanced, following research showing that this facilitates intake (Dziemianko 2014; Choi 2017). Another benefit of examples is that they can help writers better understand the use of semantically similar collocations like *INCREASING/GROWING number*.

ColloCaid also aims to provide feedback on miscollocations or collocations that sound out of place in formal academic writing. As explained earlier, at a later stage in our research we will use learner corpora to research typical problem areas that can be addressed. Preliminary data from the British Academic Written English (BAWE) corpus of university student assignments (Nesi 2011), for example, indicates novice EAP users tend to overuse informal collocations like *A LOT OF time/research/information/effort/criticism*, and brings to the surface miscollocations like **an increase OF sales/profit/interest/production*. To address this kind of problem, our preferred approach is to raise awareness and educate rather than autocorrect, as indicated in Figures 3 and 4.

Finally, we also intend to allow users to customize collocation cues according to their needs. For example, it should be possible for writers to turn off real-time help and check their texts only when they wish, and hide or restore specific collocation prompts.

[3] Strength of association based on logDice score. See Frankenberg-Garcia et al. (2018) for further details.
[4] See Frankenberg-Garcia et al. (2018) for further details.

https://doi.org/10.1017/S0261444818000277 Published online by Cambridge University Press

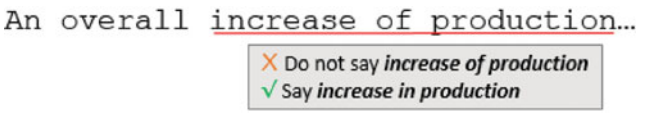

Figure 4 (Colour online) Drawing attention to miscollocations

At the time of writing this paper, we have compiled circa 50% of our target lexicographic database and are working on how to best link it to a text-editing environment. In the next steps of our research, we intend to test an initial prototype with end-users and experts in order to enhance usability and develop appropriate design solutions. Additionally, we aim to forge partnerships with researchers focusing on EAP collocation problems among specific user groups to investigate how ColloCaid can be fine-tuned to their needs.

6. Conclusion

In this paper I have argued that despite the remarkable advances that have taken place in pedagogical lexicography over the past decades, dictionaries fail to address higher-level needs of writers efficiently and are rapidly losing ground to other tools and resources. The way forward would seem to be to bring dictionary information to writers rather than to wait for writers to become better users of dictionaries, hence the necessity to investigate the integration of user needs, lexicographic data and digital writing environments. In response to this challenge, we are taking a step beyond the static dictionary through the ColloCaid project, where we are researching ways to convey information on collocation to writers as they write, with minimal disruption of the writing process. A distinguishing characteristic of ColloCaid is that it is not limited to providing feedback on miscollocations. Its main aim is to 'feed forward', raising awareness of collocations writers may not remember or know how to look up. While the prototype we are developing is specifically for EAP users, the implications of our research can be broadened to other languages and beyond academic purposes.

Acknowledgements

I would like to thank the TechLing 2017 committee for inviting me to present this plenary at the University of Bologna in Forlì. The ColloCaid project is funded by the UK Arts and Humanities Research Council Grant AH/P003508/1.

References

Ackermann, K. & Y. Chen (2013). Developing the Academic Collocations List (ACL): A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes* 12, 235–247.

Ackermann, K., J. de Jong, A. Kilgarriff & D. Tugwell (2011). *The Pearson international corpus of academic English (PICAE)*. https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-47.pdf.

https://doi.org/10.1017/S0261444818000277 Published online by Cambridge University Press

Atkins, S. & K. Varantola (1997). Monitoring dictionary use. *International Journal of Lexicography* 10.1, 1–45.
Benson, M., E. Benson & R. Ilson (1986). *The BBI dictionary of English word combinations*. Amsterdam/Philadelphia: John Benjamins.
Boers, F. & S. Webb (2017). Teaching and learning collocation in adult second and foreign language learning. *Language Teaching* 51.1, 77–89.
British National Corpus (BNC) (no date). https://corpus.byu.edu/bnc/.
Choi, S. (2017). Processing and learning of enhanced English collocations: An eye movement study. *Language Teaching Research* 21.3, 403–426.
ColloCaid (no date). http://www.collocaid.uk/.
Conklin, K. & N. Schmitt (2012). The processing of formulaic language. *Annual Review of Applied Linguistics* 32, 45–61.
Corpus of Contemporary American English (COCA) (no date). https://corpus.byu.edu/coca/.
Cowie, A. (1999). *English dictionaries for foreign learners*. Oxford: Clarendon Press.
Cowie, A. (2009). *The Oxford history of English lexicography*. Oxford: Oxford University Press.
Crossley, S., T. Salsbury & D. McNamara (2015). Assessing lexical proficiency using analytic ratings: a case for collocation accuracy. *Applied Linguistics* 36.5, 570–590.
De Schryver, G.-M. & D. Joffe (2004). On how electronic dictionaries are really used. *Proceedings of the Eleventh EURALEX*, Lorient, France, 187–196.
Durrant, P. (2016). To what extent is the Academic Vocabulary List relevant to university student writing? *English for Specific Purposes* 43, 49–61.
Durrant, P. & N. Schmitt (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching* 47.2, 157–177.
Dziemianko, A. (2014). On the presentation and placement of collocations in monolingual English learners' dictionaries: Insights into encoding and retention. *International Journal of Lexicography* 27.3, 259–279.
Dziemianko, A. (2015). Colours in online dictionaries: A case of functional labels. *International Journal of Lexicography* 28.1, 27–61.
Dziemianko, A. (2017). Dictionary form in decoding, encoding and retention: Further insights. *ReCALL* 29.3, 335–356.
Ellis, N., R. Simpson-Vlach & C. Maynard (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly* 42.3, 375–396.
Faerch, C. & G. Kasper (1983). *Strategies in interlanguage communication*. Harlow: Longman.
Flax Library (no date). http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=collocations.
Frankenberg-Garcia, A. (1999). Providing student writers with pre-text feedback. *ELT Journal* 53.2, 100–106.
Frankenberg-Garcia, A. (2005). A peek into what language learners as researchers actually do. *International Jounal of Lexicography* 18.3, 335–355.
Frankenberg-Garcia, A. (2011). Beyond L1-L2 equivalents: Where do users of English as a foreign language turn for help? *International Journal of Lexicography* 24.1, 97–123.
Frankenberg-Garcia, A. (2012a). Learners' use of corpus examples. *International Journal of Lexicography* 25.3, 273–296.
Frankenberg-Garcia, A. (2012b). Raising teachers' awareness of corpora. *Language Teaching* 45.4, 475–489.
Frankenberg-Garcia, A. (2014). The use of corpus examples for language comprehension and production. *ReCALL* 26.2, 128–146.
Frankenberg-Garcia, A. (2015). Dictionaries and encoding examples to support language production. *International Journal of Lexicography* 24.4, 490–512.
Frankenberg-Garcia, A. (2018). Investigating the collocations available to EAP writers. *Journal of English for Academic Purposes* 35, 93–104. https://doi.org/10.1016/j.jeap.2018.07.003.
Frankenberg-Garcia, A., R. Lew, J. C. Roberts, G. P. Rees & N. Sharma (2018). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL* Advance access online. doi:10.1017/S0958344018000150.
Gardner, D. & M. Davies (2014). A new Academic Vocabulary List. *Applied Linguistics* 35.3, 305–327.
Grammarly (no date). https://app.grammarly.com/.
Gromann, D. & J. Schnitzer (2016). Where do business students turn for help? An empirical study on dictionary use in foreign-language learning. *International Journal of Lexicography* 29.1, 55–99.

Hemingway (no date). http://www.hemingwayapp.com/.
Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London/NewYork: Routledge.
Hornby, A., A. Cowie & J. Lewis (1974). *Oxford advanced learner's dictionary*. London: Oxford University Press.
Hornby, A., E. Gatenby & H. Wakefield (1942). *Idiomatic and syntactic dictionary of English*. Tokyo: Kaitakusha.
Hsu, J. (2007). Lexical collocations and their relation to the online writing of Taiwanese college English majors and non-English majors. *Electronic Journal of Foreign Language Teaching* 4.2, 192–209.
Hyland, K. & P. Shaw (2016). *Introduction*. In K. Hyland & P. Shaw (eds.), *The Routledge handbook of English for academic purposes*. London: Routledge, 1–14.
Jardim, C. (2018). *Investigating the lexicographical needs of Brazilian learners of English: A user study*. Ph.D. thesis: University of Glasgow.
Just the Word (no date). http://www.just-the-word.com/.
Kilgarriff, A., V. Baisa, J. Bušta, M. Jakubíček, V. Kovvář, J Michelfeit & V. Suchomel (2014). The Sketch Engine: Ten years on. *Lexicography* 1, 7–36.
Kim, S. (2017). EFL learners' dictionary consultation behaviour during the revision process to correct collocation errors. *International Journal of Lexicography*. Advance access, doi: 10.1093/ijl/ecx009.
Kosem, I. (2010). *Designing a model for a corpus-driven dictionary of Academic English*. Ph.D. thesis: Aston University. http://publications.aston.ac.uk/14664/1/Kosem2010_484017_3.pdf .
Laufer, B. (2011). The contribution of dictionary use to the production and retention of collocations in a second language. *International Journal of Lexicography* 24.1, 29–49.
Laufer, B. & T. Waldman (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning* 61.2, 647–672.
Levy, M. & C. Steel (2015). Language learner perspectives on the functionality and use of electronic language dictionaries. *ReCALL* 27.2, 177–196.
Lew, R. (2011). Online dictionaries of English. In P. Fuertes-Olivera & H. Bergenholtz (eds.), *E-Lexicography: The internet, digital initiatives and lexicography*. London/NewYork: Continuum, 230–250.
Lew, R. (2016). Dictionaries for learners of English. *Language Teaching* 49.2, 291–294.
Lew, R. & G-M. de Schryver (2014). Dictionary users in the digital revolution. *International Journal of Lexicography* 27.4, 341–359.
Mayor, M. (2013). *Longman collocations dictionary and thesaurus*. Harlow: Pearson Education.
Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63.2, 81–97.
Müller-Spitzer, C. (2014). Empirical data on contexts of dictionary use. In C. Müller-Spitzer (ed.), *Using online dictionaries*. Berlin/Boston: Walter de Gruyter, 85–126.
Müller-Spitzer, C., S. Wolfer & A. Koplenig (2015). Observing online dictionary users: Studies using Wiktionary log files. *International Journal of Lexicography* 28.1, 1–26.
Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
Nattinger, J. & J. DeCarrico (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.
Nesi, H. (2011). BAWE: An introduction to a new resource. In A. Frankenberg-Garcia, L. Flowerdew & G. Aston (eds.), *New trends in corpora and language learning*. London: Continuum, 213–228.
Nesi, H. (2014). Dictionary use by English language learners. *Language Teaching* 47.1, 38–85.
Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam/Philadelphia: John Benjamins.
Paquot, M. (2010). *Academic vocabulary in learner writing: From extraction to analysis*. London: Continuum.
Paquot, M. (2017). L1 frequency in foreign language acquisition: Recurrent word combinations in French and Spanish EFL learner writing. *Second Language Research* 33.1, 13–32.
Paquot, M. & S. Granger (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics* 32, 130–149.
Peters, E. (2016). The lexical burden of collocations: The role of interlexical and intralexical factors. *Language Learning* 20.1, 113–138.
Procter, P. (1978). *Longman dictionary of contemporary English*. Harlow: Longman.
Ranalli, J. (2013). The online strategy instruction of integrated dictionary skills and language awareness. *Language Learning and Technology* 17.2, 75–99.
Runcie, M. (2002). *Oxford collocations dictionary for students of English*. Oxford: Oxford: Oxford University Press.

Rundell, M. (2009). *Macmillan English dictionary online*. Oxford: Macmillan Education. http://www.macmillandictionary.com/.
Rundell, M. (2010). *Macmillan collocations dictionary*. Oxford: Macmillan.
Rundell, M. (2015). From print to digital: Implications for dictionary policy and lexicographic conventions. *Lexikos* 25, 301–322.
Rundell, M. (2017). Dictionaries and crowdsourcing, wikis and user-generated content. In P. Hanks & G.-M. de Schryver (eds.), *Handbook of modern lexis and lexicography*. Berlin/Heidelberg: Springer.
Santos, D. & A. Frankenberg-Garcia (2007). The corpus, its users and their needs: A a user-oriented evaluation of COMPARA. *International Journal of Corpus Linguistics* 12.3, 335–374.
Sinclair, J. (1987a). *Collins COBUILD English dictionary for advanced learners*. London: Collins.
Sinclair, J. (1987b) (ed.). *Looking up: An account of the COBUILD project in lexical computing*. London/Glasgow: Collins ELT.
SkELL (no date). Retrieved 22 February 2018, from https://www.sketchengine.co.uk/skell/.
Sketch Engine (no date). https://www.sketchengine.co.uk.
Tarp, S., K. Fisker & P. Sepstrup (2017). L2 writing assistants and context aware dictionaries: New challenges to lexicography. *Lexikos* 27, 494–521.
Welker, H. (2006). *O Uso de Dicionários. Panorama Geral das Pesquisas Empíricas*. [Dictionary use: An overview of empirical studies]. Brasília: Thesaurus.
Wray, A. (2013). Formulaic language. *Language Teaching* 46.3, 316–334.
Write&Improve (no date). https://writeandimprove.com/.
WriteAway (no date). http://writeaway.nlpweb.org/.
Yoon, C. (2016). Concordancers and dictionaries as problem-solving tools for ESL academic writing. *Language Learning and Technology* 20.1, 209–229.
Zipf, G. (1949). *Human behavior and the principle of least effort*. Reading: Addison-Wesley.

ANA FRANKENBERG-GARCIA is Reader in Translation Studies at the University of Surrey. Her research focuses on applied uses of corpora in translation, lexicography and language learning. She was Principal Investigator of COMPARA – a 3-million-word, open-access parallel corpus of English and Portuguese fiction – and Chief Editor of the bilingual, corpus-based *Oxford Portuguese dictionary* (2015). She is currently Principal Investigator of ColloCaid, an AHRC-funded project aimed at helping writers with collocations in real time. Outside the university, Ana has led various hands-on professional development workshops on practical applications of corpora.

https://doi.org/10.1017/S0261444818000277 Published online by Cambridge University Press