

30 Using data-driven learning in language teaching

Gaëtanelle Gilquin and Sylviane Granger

1. The pedagogical functions of data-driven learning

Data-driven learning (DDL) consists of using the tools and techniques of corpus linguistics for pedagogical purposes (see Chapter 29, this volume). After a modest beginning in the 1980s, the approach started to attract more advocates and the number of publications began to rise, especially after the turn of the millennium, which can presumably be related to the spread of technologies in language teaching. This rise in DDL publications goes hand in hand with diversification of DDL practices. For example, while DDL was for a long time confined to English, studies on other languages have begun to appear (e.g. Yao 2019), while it was initially more oriented towards university students, its use with young learners has become more widespread (e.g. Crosthwaite 2020), and while it was first almost exclusively limited to writing, activities involving spoken and multimodal corpora have now emerged (e.g. Hirata 2020; Meunier 2020). Recent meta-analyses have started to take stock of the body of research on DDL, either in general (e.g. Boulton and Cobb 2017; Pérez-Paredes 2019) or with a more specific focus (e.g. Lee et al. 2019 on DDL for vocabulary learning; Mizumoto and Chujo 2015 on DDL for English learning in Japan).

The use of DDL in language teaching presents several advantages. The first one is that it brings authenticity into the classroom. Not only do corpora make it possible to expose learners to authentic language, but they can actually present them with a large number of authentic instances of a particular linguistic item. This ‘condensed exposure’ (Gabrielatos 2005: 10) can, among others, contribute to vocabulary expansion or heightened awareness of language patterns.

Secondly, DDL has an important corrective function. Learners, by comparing their own writing with data produced by (native) expert writers or by consulting a learner corpus where errors have been annotated, can find the help they need to correct their own interlanguage features (misuse, overuse, underuse) and thus improve their writing. As pointed out by Nesselhauf (2004: 140), this is particularly useful for fossilised errors.

The DDL approach also has the advantage of including an element of discovery which arguably makes learning more motivating and more fun. In the DDL literature, learners are described, alternatively, as travellers (Bernardini 2001: 227), researchers (Johns 1997: 101) or detectives with Johns’s (1997: 101) slogan ‘Every student a Sherlock Holmes’. By means of various activities (see Section 3), learners are encouraged to observe corpus data, make hypotheses and formulate rules in order to gain insights into language (inductive approach), or check the validity of rules from their grammars or textbooks (deductive approach). They thus become more involved, more active and, ultimately, more autonomous in the learning process.

More generally, learners are said to acquire (or at least refine) a number of crucial learning skills through the use of DDL. O’Sullivan (2007: 277) lists the following: ‘predicting, observing, noticing, thinking, reasoning, analysing, interpreting, reflecting, exploring, making inferences (inductively or deductively), focusing, guessing, comparing, differentiating, theorising, hypothesising, and verifying’. These skills can be used to explore language, but since they are general cognitive skills, they may also be transferred to other fields of study.

2. Data-driven learning material

In order to adopt a DDL methodology, two main resources are needed: a corpus and a tool to exploit it. The choice of the corpus is crucial. It would probably not be wrong to say that any type of corpus may be used in DDL, and indeed, the literature on DDL mentions quite a large range of corpora: written, spoken or multimodal, monolingual or bilingual, general or specialised, native or non-native, raw or annotated, etc. As can be expected, however, particular corpora are best suited for certain purposes. English for Specific Purposes corpora, for example, allow teachers to ‘equip students with the communicative skills they need to participate in particular disciplinary contexts’ (Hyland 2019: xiii; see Chapters 24 and 28, this volume). Bilingual corpora, which show original texts and their translations side by side, may help translation trainees by ‘drawing [their] attention to (un)typical solutions for typical problems found by *mature*, expert translators’ (Bernardini 2004: 20, emphasis original; see also Chapter 34, this volume). Learner corpora, which are made up of data produced by non-native speakers

of the language, can be extremely useful for form-focused instruction (see, e.g., Granger and Tribble 1998) by making students aware of typical interlanguage features, while also contributing to increased motivation.

Whatever the type of corpus chosen, one important issue is that of its authenticity. Of course, in a way, corpora are always authentic in the sense that they contain language produced in natural settings. However, scholars like Widdowson (2000) argue that corpora may lack authenticity at the receptive end. Thus, learners may find it hard to relate to texts that were produced in a different culture, within a context that is not necessarily familiar to them. Sripicharn (2004), who compares how native speakers and learners deal with concordance-based tasks, shows that native speakers are better able than learners to contextualise the concordance lines (by identifying the setting of the concordances and the text type from which they are extracted). Several solutions have been proposed in the literature to help learners ‘authenticate’ (Widdowson 2003: 66) the materials they are working with, for example using a corpus of recent news items (Chambers 2005) or asking the students to compile their own ‘personal corpus’ (Chen and Flowerdew 2018: 343), which could for instance include research articles in their discipline (see Chapters 24 and 29, this volume).

Two types of corpora that appear particularly helpful for the process of authentication are the pedagogic corpus (Willis 2003) and the local learner corpus (Seidlhofer 2002). The former is a corpus ‘created not for linguistic research but specifically for language learning purposes’ (Chambers 2019: 464), which could for example consist of the texts

used in the classroom to support teaching. Such corpora are better contextualised and more directly relevant to learners.

The second type of corpus that seems promising in terms of authenticity is the local learner corpus, which contains data produced by the very same students who use the corpus. These students are thus ‘both participants in and analysts of their own language use’ (Seidlhofer 2002: 213), and the interlanguage features represented in the corpus are the features of their own interlanguage. This also means that the teacher can provide ‘tailor-made feedback’ (Mukherjee 2006: 19) to the learners, either as a group or individually. As shown by Cotos (2014: 218), the use of local learner corpora is a particularly effective practice that can be readily integrated into DDL instruction.

Another important issue when it comes to the choice of a corpus is annotation (see Chapter 2, this volume). Corpora can be used as raw text, i.e. with no annotation of any kind, or they can be annotated with additional information such as part-of-speech (POS-tagging), syntactic structure (parsing) or, in the case of learner corpora, errors (error-tagging). Raw corpora offer numerous possibilities for the exploration of language by learners. However, they also have their limitations. For example, a raw corpus may involve a lot of editing (by the teacher or learner) to get rid of unwanted concordance lines, whereas a POS-tagged or parsed corpus may help refine the search query (for example, selecting *to* as a preposition and not as an infinitive marker) and thus reduce the amount of necessary editing. As for error-tagging, it makes it much easier to notice interlanguage features and often comes with possible corrections, which can help learners revise their own texts.

Corpora are of little help if they are not combined with a tool to exploit them (see Chapters 9 and 10, this volume). Corpus query tools used in DDL should be ‘learner-friendly’ (Lee et al. 2019: 747) and accessible to users who have limited experience with corpora. As rightly underlined by Crosthwaite and Cheung (2019: 171), ‘[c]omplex or unappealing corpus query tools are (...) one of the biggest hurdles for the successful uptake of DDL’. Kilgarriff et al. (2015: 66), for instance, describe how they designed *SKELL* (*Sketch Engine for Language Learners*) as ‘a stripped-down, non-scary version of Sketch Engine for use by learners’. *BNClab* (Brezina et al. 2018), a platform developed at Lancaster University, makes it easy to search a spoken sample of the *British National Corpus* and to compare the first edition of the corpus with the 2014 edition. The platform is particularly attractive through data visualisation, which arguably makes learners more receptive to the information they discover. *TLC Hub*, also developed at Lancaster University, offers the same facilities to query learner speech from the *Trinity Lancaster Corpus* (Gablasova et al. 2019; <http://cass.lancs.ac.uk/trinity-lancaster-corpus>). Other tools, like the MICASE and MICUSP interfaces or Mark Davies’s English-Corpora.org website, can also be useful for learners, especially if they are combined with some guides explaining how to use the tools (see, e.g., Poole 2018 for English-Corpora.org). *WordSmith Tools* (Scott 2020) has an interesting option for DDL, namely the possibility of creating exercises by blanking out the search words in concordances.

3. The operationalisation of data-driven learning

The range of activities that are possible in DDL is wide and, as Breyer (2006: 162) puts it, ‘limited only by the imagination of the user’. Space prevents the reviewing of all these possibilities, but in this section, we give a broad overview of the way DDL may be operationalised and we show how the choice of presentation and activity may depend on a variety of factors (see Chapters 28, 29 and 31, this volume).

The first major decision when doing DDL is whether to let learners query the corpus themselves (direct approach) or to provide them with materials prepared by the teacher on the basis of corpus analysis (indirect approach) (see Yoon and Jo 2014 on this distinction). This choice depends on the availability or not of the necessary hardware and software for learners, but it is also a function of learners’ level. Thus, Boulton (2008a: 38) suggests that ‘DDL in early stages can eliminate the computer from the equation by using prepared materials on paper’. The effect of a direct vs indirect approach to DDL is still relatively unclear. While it has often been claimed that the indirect approach, through the use of printed materials, is more effective, Boulton and Cobb’s (2017) meta-analysis shows that the opposite seems to be true and Vyatkina (2016) finds the two approaches equally effective.

In indirect DDL, teachers have the possibility of presenting the corpus data in various ways. Concordance lines may be truncated (KWIC – Key-Word-In-Context – view) or take the shape of complete sentences, the whole concordance may be provided to the learner or just a selection of it, and the concordance lines may be edited or presented in

their original form. Each presentation has its pros and cons. The KWIC view may be confusing, especially for beginners. Johns (1986: 157) observes that learners' first reaction is often to complain about the 'unfinished sentences'. However, the KWIC view, with all the occurrences of the search word aligned under one another, makes patterns more visible than the sentence view, and Boulton (2009a) actually reports an experiment with lower-intermediate learners where KWICs provided better results than full sentence contexts. Next, the concordancer output may be used as is or be manipulated in some way. Manipulation may involve the selection of a subset of the concordance, often with the aim of reducing the data to manageable quantities. Several criteria have been proposed in the literature to perform this selection, including readability (the most difficult concordance lines are discarded, cf. Kuo et al. 2001), frequency (only the concordance lines illustrating the most frequent uses are kept, cf. Levy 1990) and usefulness (only those concordance lines that are judged useful are kept, cf. Tribble 1997). DDL has been claimed to be more effective when the concordance lines are carefully selected by the teacher (Lee et al. 2019). However, as rightly pointed out by Gabrielatos (2005: 18), '[t]his manipulation should be carried out with the understanding that the adapted samples are not good guides to the frequency of a language item'. Manipulation may also consist in editing the concordances, either superficially (e.g. through colour coding or highlighting, cf. Wicher 2020) or, more substantially, by simplifying them (cf. Gabrielatos 2005). While such simplification might be said to undermine the authenticity advantage of DDL and leave learners unprepared 'for the realities of the authentic language we are presumably preparing them for' (Boulton 2009b: 89), Nesselhauf's (2004) discussion of a concordance of the verb *suggest* in the LOCNESS corpus (a corpus of argumentative essays by British and

American students) and the German subcorpus of the *International Corpus of Learner English* is a good example of why manipulation is sometimes necessary in DDL (especially with beginners):

As the lines are now, they could be confusing for learners in many respects: there is at least one typographical error (*suggested than* instead of *that* in LOCNESS); one of the occurrences of *to* after *suggest* does not constitute wrong complementation (*could suggest to her two colleagues*); and, as *suggest + ing* only occurs in the learner but not in the native speaker corpus, learners might even come to the conclusion that this construction is not possible in English.

(Nesselhauf 2004: 143-4)

DDL activities may be located along a cline ranging from teacher-led to learner-led (Mukherjee 2006; see also Gabrielatos 2005). At the teacher-led end, we find relatively controlled tasks such as cloze tests and fill-in exercises. At the learner-led end, we find what Bernardini (2004: 22) calls ‘discovery learning’, which involves ‘brows[ing] large and varied text collections in open-ended, exploratory ways’. As we move from one end of the cline to the other, learners have more freedom and bear more responsibility for their own instruction, deciding, for example, what they are going to investigate and how they want to go about it. In-between totally teacher-led DDL and totally learner-led DDL, there is a whole range of activities, with various types of ‘filter’ exercised by the teacher (Gavioli 2005: 30). By way of illustration, here are a few activities that could be undertaken in the classroom. Learners could be shown a concordance sorted alphabetically and be encouraged to notice the repetition of certain lexical chunks or be

asked to group the patterns in a meaningful way. They could be given a series of blanked out concordances illustrating different contexts of a word and be required to find the missing word. Alternatively, the concordances could come from a bilingual corpus, so that learners can use the translations to help them find the missing word. The potential of DDL for editing one's own work has already been mentioned earlier, and many studies in the literature show how students can use corpora to revise their work, either by correcting problems underlined by the teacher or by deciding themselves what they want to check with the help of the corpus (Kennedy and Miceli 2001; see also Crosthwaite and Stell 2020 on DDL for text revision in a first language context). One use that tends to be neglected in the DDL literature is the use of the corpus as a 'sleeping resource' (Johns 1988: 22), to help learners when the need arises, for example to answer a question that has suddenly arisen during the lesson.

While concordances are probably the main component of DDL, other formats are possible. Frequency lists, which list all the words of the corpus in descending order of frequency, may prove to be a valuable resource as well, for example as a means of learning more about the subject of a text and its meaning (Aston 2001). It is also possible to compare two frequency lists built on the basis of different corpora. The comparison could involve two varieties of English (e.g. British vs American English) or two genres or text types (e.g. fiction vs journalese, speech vs writing), with the aim of making learners more sensitive to language variation. Learners could also compare a frequency list representing learner production (ideally, their own production) and one representing expert production, which would make them aware of the words that tend to be underused or overused by learners. The words from the frequency lists can then be

used as a starting point for concordance analysis. Most corpus query tools nowadays also enable users to automatically extract collocates and lexical bundles. These, like frequency lists, may serve as a good starting point for the further analysis of language – in this case, language in its phraseological dimension. Another way to exploit corpora in the classroom is to read entire portions of the corpus. This is what Charles (2007: 295) recommends for the study of rhetorical functions: a particular search item is used as ‘a probe to locate the part of a text in which a given rhetorical function may occur’, and the context is then expanded to the whole paragraph or the whole text to see how the function is expressed. Finally, although DDL has mainly been used with written data, it can also be used with speech, provided spoken or multimodal corpora are available. Meunier (2020: 13) hence recommends ‘adopting new ways of DDL-izing teaching practices by including more multimodality (sound, video, pictures)’ and provides illustrations of how internet videos can be used for DDL.

The choice of presentation and activity depends on a number of factors, among which the learning context, learners’ level and the phenomenon investigated. Thus, the direct approach to DDL and focus on the learner’s individual needs may be easier in the context of a one-to-one consultation (such as described by Johns 2002) than in the context of a classroom activity. An activity meant to be carried out as part of a homework assignment or in distance education (cf. Collins 2000) may have to contain more explicit instructions and ‘signposts’ than one which takes place in class, with the teacher as a guide and the fellow students as ‘travel companions’. The learner’s level (both in terms of language proficiency and familiarity with DDL) is also important when deciding on a particular presentation of the data and a specific activity. This

probably explains why scholars do not seem to agree on the level of the learners for whom DDL is appropriate. Depending on the methods they have in mind, DDL may be suitable for a given audience or not. In fact, it has been argued that DDL is possible with all learners (even beginners), but that (i) the method has to be adapted to the learners' level (Hadley 2002) and (ii) results may vary, with beginners 'draw[ing] relatively low-level conclusions about the structuring of the language' and more advanced learners 'mak[ing] more subtle high-level inferences' on the basis of one and the same concordance (Johns 1986: 159). Meta-analyses like Boulton and Cobb (2017) and Lee et al. (2019) however suggest that DDL works best from an intermediate level upwards. Another aspect that should be taken into account when choosing a way of approaching DDL is the phenomenon to be investigated. Many authors claim that DDL is most effective 'on the "collocational border" between syntax and lexis' (Johns 2002: 109; see also Levy 1990 or O'Sullivan and Chambers 2006). However, it can also work with other phenomena and with larger units than individual words, for example syntactic constructions (e.g. Gilquin 2021) or aspects of discourse (e.g. Charles 2007), although, here again, some adaptation may be required.

A final note is that, despite the attraction of some methods (like concordances) and some phenomena (Kennedy and Miceli 2001: 83 refer to the 'lure of prepositions'), a key word in DDL is variety. Not only does variety make it possible to prevent tediousness among learners (a problem often highlighted in the literature, cf. Chambers 2007), but it also caters to learners' different preferences. Similarly, since it is precisely one of the goals of DDL to foster more autonomous learning, the teacher should avoid

conformity as far as possible and agree to let the students approach corpora in the way they feel most comfortable with (Hunston 2002).

4. Evaluating data-driven learning

Like any pedagogical method, it is important that DDL should be properly evaluated.

As pointed out by Boulton (2008b), three types of evaluation can be carried out:

evaluation of the attitudes (what do participants think about DDL?), practices (how well are learners doing with DDL?) and efficiency (can learners gain benefit from DDL?).

While attitudes and practices are important, the criterion that, ultimately, should be decisive in determining whether DDL is worth doing or not is efficiency. Learners may enjoy DDL and be good at it, but if they do not learn anything from it, there is no point in adding this to a curriculum that is already overloaded.

As regards attitudes, a survey of the literature reveals generally positive feelings among learners (e.g. Mizumoto and Chujo 2015), but also some negative ones (e.g. Quan 2016). Certain authors even report both positive and negative attitudes among the same learners. The students in Kennedy and Miceli (2001), for example, found DDL helpful and confidence-boosting, but sometimes also discouraging, time-consuming and frustrating. What is often neglected in such evaluative studies and should be considered more closely is teachers' attitudes towards DDL. The evaluation of practices, i.e. learners' capacities to do DDL, also produces mixed results (cf. Hadley 2002; Kennedy and Miceli 2001). As for efficiency, which can be considered crucial for the future of

DDL, recent meta-analyses like Boulton and Cobb (2017) reveal generally large effect sizes for comparisons of pre- and post-tests or control and experimental groups, which suggests that DDL does work. However, some caveats should be mentioned. First, effect sizes may display a great deal of variation according to the moderator variables that are taken into account in these meta-analyses (study design, language focus, etc.). Second, these meta-analyses only consider certain (types of) studies – in particular, experimental and quantitative studies. For approaches that are not (or not well) represented in the meta-analyses, the jury is still out. Thus, although Boulton and Cobb (2017: 378) mention rather positive results for the few studies that have evaluated the use of corpora as reference resources, Ackerley (2017) obtains mixed results in a free writing task, while observing that short-answer activities such as gap-filling tend to generate more positive results. Third, delayed post-testing is often lacking (Boulton and Cobb 2017), which means that the longer-term effects on the development of more general skills that is regularly presented as a strength of DDL have so far not really been evaluated. Given the difficulty of assessing such effects (see Cobb and Boulton 2015), however, this is unlikely to change anytime soon. As this brief survey shows, many results are still inconclusive and even contradictory. The effects of DDL are dependent on so many different variables that replication studies are particularly needed in this field and should be encouraged.

5. The problems and limitations of data-driven learning

While the evaluations of DDL seem to suggest that its future ‘looks rather bright’ (Boulton and Cobb 2017: 388), one must recognise along with Mizumoto and Chujo (2015: 12) that ‘not so many researchers and practitioners make use of its potential in their teaching practice’. This may be because the implementation of DDL in the classroom presents a number of difficulties, some of which are examined here.

Logistics has often been cited as one of the biggest problems of DDL. Thanks to the increasingly widespread use of laptops among students, as well as the development of DDL on mobile devices (cf. Quan 2016), in most situations this is no longer seen as a major problem (although it may still be a real issue in less developed countries). In addition, many resources are now freely available: corpus query tools like *AntConc* (Anthony 2019) and a large number of texts or off-the-shelf corpora downloadable from the Internet (see Chapter 9, this volume). What is still largely missing is ready-made DDL materials for teachers, in the form of worksheets, teaching plans or textbooks. For teachers who have no such materials available to them, the preparation of DDL activities is time-consuming. Equally time-consuming is the DDL task itself (Díez Bedmar 2006 describes an activity that took one hour and a half for just one word). And according to some, the results ‘might not repay the time taken’ (Willis, cited in Hunston 2002: 178).

If we consider the teacher’s point of view, one reason for not doing DDL might simply be that the teacher does not know enough about corpora and the possibility of using

corpora in the classroom. There is therefore a need for better integration of DDL into teacher training programmes (Hirata 2020; see also Chapter 32, this volume), as well as guides to using corpora that are specifically designed for teachers (e.g. Friginal 2018). As Mauranen (2004: 100) points out, '[b]efore learners can be introduced to good corpus skills, their teachers need to possess them in the first place'. In some cases, the teacher may be familiar with corpora and DDL but prefer, for some reason, not to adopt such a methodology with his/her students. A number of practical objections might be raised by the teacher, but these may actually be 'camouflage for more profound theoretical concerns about the nature of learning' (Boulton 2009b: 99), and more especially about the teacher's possible loss of control over what happens during the lesson, which may be experienced as a 'loss of expertise' by the teacher (Hunston 2002: 171).

For learners too, DDL may sometimes appear rather off-putting. Although most of them will be used to technologies, working with corpora may not be straightforward to them and may necessitate quite some training to acquire the basic skills (reading concordances, interpreting lists of collocates, etc.) – what Mukherjee (2002: 179) calls 'corpus literacy' (see also Sripicharn 2010 on the importance of preparing learners for using corpora). Students may have 'difficulty devising effective search strategies' (O'Sullivan and Chambers 2006: 60) because of faulty spelling, for example, or simply because of the complexity of the processes involved (see Sun 2003 for a good example of ineffective search strategy), and they may draw wrong inferences on the basis of the evidence – what O'Keeffe (2021) calls 'fake discovery' (see Sripicharn 2004 for an illustration). It must also be stressed that DDL (and, in particular, the inductive learning

strategies that it often entails) may be more suitable for certain learners than others (Gilquin 2021).

The final aspect that may explain some of the reluctance to apply DDL has to do with its content. We have already mentioned the problem of authentication (or lack thereof) of the corpus data. In addition, the output of the search query may contain too much or too little data, or no data at all – which, for learners, poses the question of the distinction between ‘The phenomenon does not exist’ and ‘There is no evidence for it in this corpus’ (Kennedy and Miceli 2001: 86). It may include too much noise, i.e. irrelevant hits, or it may be too difficult for learners to understand because of insufficient knowledge of the target language (Koosha and Jafarpour 2006: 206). It could also be that the corpus shows more details than the student is expected to learn or contains language which the teacher does not want the students to imitate, for example non-standard forms, swear words or literary phrases. Another problem is that the DDL approach may not be effective for all aspects of language, with some questions more ‘concordance-ready’ than others (Johns 1988: 25). As pointed out by Ädel (2010: 48), DDL has tended to focus on ‘surface forms’, especially lexico-grammatical features, which lend themselves well to corpus analysis, while disregarding genre-related aspects which are less obviously tractable (but see Cotos et al. 2017; Crosthwaite and Cheung 2019). More generally, Gabrielatos (2005: 21) warns against the dangers of ‘corpus worship’ and ‘frequency worship’, hence suggesting that corpora are no panacea and that other factors besides frequency have to be considered for teaching to be successful.

These problems should not deter us from doing DDL, which, as was shown above, is a promising technique bringing learners into contact with potentially large amounts of authentic language, motivating them through the element of discovery, developing important cognitive skills and, more generally, providing benefits which arguably go well beyond the knowledge of the phenomenon under study and could really empower students. However, it is important to use this technique in full awareness of its limitations to avoid the risk of ‘DDL worship’. This should imply, among other things, adapting DDL to the specific learning situation in which it takes place, but also better integrating DDL into the national curriculum (Papaioannou et al. 2020) and combining it with more traditional teaching methods (Chen and Flowerdew 2018). If DDL is to become a viable technique, it is also crucial that it should move out of its favourite spheres of action: from English classes to classes in other languages and from universities to secondary and primary education (see Vyatkina 2020), but also from the classroom to everyday life (see Chen and Flowerdew 2018; Meunier 2020). Studies have shown that learners usually stop using corpora after the in-class DDL training/activities (see, e.g., Crosthwaite and Cheung 2019), although, as shown by Charles (2014), when learners build their personal corpus in their own field, they tend to continue using it after the course has finished. Capitalising on the young generation’s technological habits (cf. the parallel drawn between corpus consultation and internet searches in Cobb and Boulton 2015), one needs to induce a transfer of corpus skills from a language learning situation to an everyday language use situation, thus turning ‘data-driven learning’ into ‘data-driven use’ and making language *learners* evolve into autonomous language *users*.

Further reading

Boulton, A. and Cobb, T. (2017) 'Corpus Use in Language Learning: A Meta-analysis', *Language Learning* 67(2): 348-93. (This meta-analysis considers the learning outcomes of DDL in 64 (quasi-)experimental studies and attempts to identify relevant moderator variables.)

Chen, M. and Flowerdew, J. (2018) 'A Critical Review of Research and Practice in Data-Driven Learning (DDL) in the Academic Writing Classroom', *International Journal of Corpus Linguistics* 23(3): 335-69. (Focusing on the context of academic writing, this review paper discusses 37 empirical studies and offers recommendations for future DDL.)

Leńko-Szymańska, A. and Boulton, A. (eds) (2015) *Multiple Affordances of Language Corpora for Data-Driven Learning*, Amsterdam: John Benjamins. (This volume presents advances in DDL, based on corpora for language learning, skills development and translation training, and using a variety of research methodologies.)

O'Keeffe, A. (2021) 'Data-Driven Learning – A Call for a Broader Research Gaze', *Language Teaching* 54(2): 259-72. (This paper provides an innovative perspective by showing how DDL can be positioned theoretically. It also identifies avenues for further research in DDL.)

Poole, R. (2018) *A Guide to Using Corpora for English Language Learners*, Edinburgh: Edinburgh University Press. (Designed for both language learners and language teachers, this book provides illustrated examples of how to search corpora, interpret findings and build one's own corpus; it includes many screenshots and hands-on exercises.)

References

Ackerley, K. (2017) 'Effects of Corpus-Based Instruction on Phraseology in Learner English', *Language Learning and Technology* 21(3): 195-216.

Ädel, A. (2010) 'Using Corpora to Teach Academic Writing: Challenges for the Direct Approach', in M. C. Campoy-Cubillo, B. Bellés-Fortuño and M. L. Gea-Valor (eds) *Corpus-Based Approaches to English Language Teaching*, London: Continuum, pp. 39-55.

Anthony, L. (2019) AntConc (Version 3.5.8) [Computer Software]. Tokyo: Waseda University. Available from <https://www.laurenceanthony.net/software>.

Aston, G. (2001) 'Learning with Corpora: An Overview', in G. Aston (ed.) *Learning with Corpora*, Houston: Athelstan, pp. 7-45.

Bernardini, S. (2001) “‘Spoilt for Choice’: A Learner Explores General Language Corpora”, in G. Aston (ed.) *Learning with Corpora*, Houston: Athelstan, pp. 220-49.

Bernardini, S. (2004) ‘Corpora in the Classroom. An Overview and some Reflections on Future Developments’, in J. Sinclair (ed.) *How to Use Corpora in Language Teaching*, Amsterdam: John Benjamins, pp. 15-36.

Boulton, A. (2008a) ‘DDL: Reaching the Parts Other Teaching Can’t Reach?’, in A. Frankenberg-Garcia (ed.) *Proceedings of the 8th Teaching and Language Corpora Conference*, Lisbon: Associação de Estudos e de Investigação Científica do ISLA-Lisboa, pp. 38-44.

Boulton, A. (2008b) ‘Esprit de Corpus : Promouvoir l’Exploitation de Corpus en Apprentissage des Langues’, *Texte et Corpus* 3: 37-46.

Boulton, A. (2009a) ‘Testing the Limits of Data-Driven Learning: Language Proficiency and Training’, *ReCALL* 21(1): 37-54.

Boulton, A. (2009b) ‘Data-Driven Learning: Reasonable Fears and Rational Reassurance’, *Indian Journal of Applied Linguistics* 35(1): 81-106.

Boulton, A. and Cobb, T. (2017) ‘Corpus Use in Language Learning: A Meta-Analysis’, *Language Learning* 67(2): 348-93.

Breyer, Y. (2006) 'My Concordancer: Tailor-Made Software for Language Teachers and Learners', in S. Braun, K. Kohn and J. Mukherjee (eds) *Corpus Technology and Language Pedagogy*, Frankfurt: Peter Lang, pp. 157-76.

Brezina, V., Gablasova, D. and Reichelt, S. (2018) BNCLab.

<http://corpora.lancs.ac.uk/bnclab> [electronic resource], Lancaster University.

Chambers, A. (2005) 'Integrating Corpus Consultation in Language Studies', *Language Learning and Technology* 9(2): 111-25.

Chambers, A. (2007) 'Popularising Corpus Consultation by Language Learners and Teachers', in E. Hidalgo, L. Quereda and J. Santana (eds) *Corpora in the Foreign Language Classroom*, Amsterdam: Rodopi, pp. 3-16.

Chambers, A. (2019) 'Towards the Corpus Revolution? Bridging the Research-Practice Gap', *Language Teaching* 52(4): 460-75.

Charles, M. (2007) 'Reconciling Top-Down and Bottom-Up Approaches to Graduate Writing: Using a Corpus to Teach Rhetorical Functions', *Journal of English for Academic Purposes* 6: 289-302.

Charles, M. (2014) 'Getting the Corpus Habit: EAP Students' Long-Term Use of Personal Corpora', *English for Specific Purposes* 35: 30-40.

Chen, M. and Flowerdew, J. (2018) 'A Critical Review of Research and Practice in Data-Driven Learning (DDL) in the Academic Writing Classroom', *International Journal of Corpus Linguistics* 23(3): 335-69.

Cobb, T. and Boulton, A. (2015) 'Classroom Applications of Corpus Analysis', in D. Biber and R. Reppen (eds) *Cambridge Handbook of English Corpus Linguistics*, Cambridge: Cambridge University Press, pp. 478-97.

Collins, H. (2000) 'Materials Design and Language Corpora: A Report in the Context of Distance Education', in L. Burnard and T. McEnery (eds) *Rethinking Language Pedagogy from a Corpus Perspective*, Frankfurt: Peter Lang, pp. 51-63.

Cotos, E. (2014) 'Enhancing Writing Pedagogy with Learner Corpus Data', *ReCALL* 26(2): 202-24.

Cotos, E., Link, S. and Huffman, S. (2017) 'Effects of DDL Technology on Genre Learning', *Language Learning and Technology* 21(3): 104-30.

Crosthwaite, P. (ed.) (2020) *Data-Driven Learning for the Next Generation: Corpora and DDL for Pre-Tertiary Learners*, London: Routledge.

Crosthwaite, P. and Cheung, L. (2019) *Learning the Language of Dentistry: Disciplinary Corpora in the Teaching of English for Specific Academic Purposes*, Amsterdam: John Benjamins.

Crosthwaite, P. and Stell, A. (2020) “‘It Helps me Get Ideas on How to Use my Words’: Primary School Students’ Initial Reactions to Corpus Use in a Private Tutoring Setting’, in P. Crosthwaite (ed.) *Data-Driven Learning for the Next Generation: Corpora and DDL for Pre-Tertiary Learners*, London: Routledge, pp. 150-70.

Díez Bedmar, M. B. (2006) ‘Making Friends with DDL: Helping Students Enrich their Vocabulary’, *Humanising Language Teaching* 8(3).

Friginal, E. (2018) *Corpus Linguistics for English Teachers. New Tools, Online Resources, and Classroom Activities*, New York: Routledge.

Gablasova, D., Brezina, V. and McEnery, T. (2019) ‘The Trinity Lancaster Corpus: Development, Description and Application’, *International Journal of Learner Corpus Research* 5(2): 126-58.

Gabrielatos, C. (2005) ‘Corpora and Language Teaching: Just a Fling or Wedding Bells?’, *TESL-EJ* 8(4): 1-35.

Gavioli, L. (2005) *Exploring Corpora for ESP Learning*, Amsterdam: John Benjamins.

Gilquin, G. (2021) ‘Using Corpora to Foster L2 Construction Learning: A Data-Driven Learning Experiment’, *International Journal of Applied Linguistics* 31(2): 229-47.

- Granger, S. and Tribble, C. (1998) 'Learner Corpus Data in the Foreign Language Classroom: Form-Focused Instruction and Data-Driven Learning', in S. Granger (ed.) *Learner English on Computer*, London: Longman, pp. 199-209.
- Hadley, G. (2002) 'An Introduction to Data-Driven Learning', *RELC Journal* 33(2): 99-124.
- Hirata, E. (2020) 'The Development of a Multimodal Corpus Tool for Young EFL Learners: A Case Study on the Integration of DDL in Teacher Education', in P. Crosthwaite (ed.) *Data-Driven Learning for the Next Generation: Corpora and DDL for Pre-Tertiary Learners*, London: Routledge, pp. 88-105.
- Hunston, S. (2002) *Corpora in Applied Linguistics*, Cambridge: Cambridge University Press.
- Hyland, K. (2019) 'Foreword: Corpora and Specialised English in the University Curriculum', in P. Crosthwaite and L. Cheung (eds) *Learning the Language of Dentistry: Disciplinary Corpora in the Teaching of English for Specific Academic Purposes*, Amsterdam: John Benjamins, pp. xi-xiv.
- Johns, T. (1986) 'Micro-Concord: A Language Learner's Research Tool', *System* 14(2): 151-62.

Johns, T. (1988) 'Whence and Whither Classroom Concordancing?', in T. Bongaerts, P. de Haan, S. Lobbe and H. Wekker (eds) *Computer Applications in Language Learning*, Dordrecht: Foris, pp. 9-33.

Johns, T. (1997) 'Contexts: The Background, Development and Trialling of a Concordance-Based CALL Program', in A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds) *Teaching and Language Corpora*, London: Longman, pp. 100-15.

Johns, T. (2002) 'Data-Driven Learning: The Perpetual Challenge', in B. Kettemann and G. Marko (eds) *Teaching and Learning by Doing Corpus Analysis*, Amsterdam: Rodopi, pp. 107-17.

Kennedy, C. and Miceli, T. (2001) 'An Evaluation of Intermediate Students' Approaches to Corpus Investigation', *Language Learning and Technology* 5(3): 77-90.

Kilgarrriff, A., Marcowitz, F., Smith, S. and Thomas, J. (2015) 'Corpora and Language Learning with the Sketch Engine and SkELL', *Revue Française de Linguistique Appliquée* XX(1): 61-80.

Koosha, M. and Jafarpour, A. A. (2006) 'Data-Driven Learning and Teaching Collocation of Prepositions: The Case of Iranian EFL Adult Learners', *Asian EFL Journal* 8(4): 192-209.

Kuo, C.-H., Wible, D., Wang, C.-C. and Chien, F. (2001) 'The Design of a Lexical Difficulty Filter for Language Learning on the Internet', in *Proceedings of the IEEE International Conference on Advanced Learning Techniques (ICALT'01), Madison, WI, 6-8 August 2001*, pp. 53-4.

Lee, H., Warschauer, M. and Lee, J. H. (2019) 'The Effects of Corpus Use on Second Language Vocabulary Learning: A Multilevel Meta-Analysis', *Applied Linguistics* 40(5): 721-53.

Levy, M. (1990) 'Concordances and their Integration into a Word-Processing Environment for Language Learners', *System* 18(2): 177-88.

Mauranen, A. (2004) 'Spoken Corpus for an Ordinary Learner', in J. Sinclair (ed.) *How to Use Corpora in Language Teaching*, Amsterdam: John Benjamins, pp. 89-105.

Meunier, F. (2020) 'A Case for Constructive Alignment in DDL: Rethinking Outcomes, Practices, and Assessment in (Data-Driven) Language Learning', in P. Crosthwaite (ed.) *Data-Driven Learning for the Next Generation: Corpora and DDL for Pre-Tertiary Learners*, London: Routledge, pp. 13-30.

Mizumoto, A. and Chujo, K. (2015) 'A Meta-Analysis of Data-Driven Learning Approach in the Japanese EFL Classroom', *English Corpus Studies* 22: 1-18.

Mukherjee, J. (2002) *Korpuslinguistik und Englischunterricht: Eine Einführung*, Frankfurt: Peter Lang.

Mukherjee, J. (2006) 'Corpus Linguistics and Language Pedagogy: The State of the Art - and Beyond', in S. Braun, K. Kohn and J. Mukherjee (eds) *Corpus Technology and Language Pedagogy*, Frankfurt: Peter Lang, pp. 5-24.

Nesselhauf, N. (2004) 'Learner Corpora and their Potential for Language Teaching', in J. Sinclair (ed.) *How to Use Corpora in Language Teaching*, Amsterdam: John Benjamins, pp. 125-52.

O'Keeffe, A. (2021) 'Data-Driven Learning – A Call for a Broader Research Gaze', *Language Teaching* 54(2): 259-72.

O'Sullivan, Í. (2007) 'Enhancing a Process-Oriented Approach to Literacy and Language Learning: The Role of Corpus Consultation Literacy', *ReCALL* 19(3): 269-86.

O'Sullivan, Í. and Chambers, A. (2006) 'Learners' Writing Skills in French: Corpus Consultation and Learner Evaluation', *Journal of Second Language Writing* 15: 49-68.

Papaioannou, V., Mattheoudakis, M. and Agathopoulou, E. (2020) 'Data-Driven Learning in a Greek Secondary Education Setting: The Implementation of a Blended

Approach’, in P. Crosthwaite (ed.) *Data-Driven Learning for the Next Generation: Corpora and DDL for Pre-Tertiary Learners*, London: Routledge, pp. 187-207.

Pérez-Paredes, P. (2019) ‘A Systematic Review of the Uses and Spread of Corpora and Data-Driven Learning in CALL Research during 2011–2015’, *Computer Assisted Language Learning*, DOI: 10.1080/09588221.2019.1667832.

Poole, R. (2018) *A Guide to Using Corpora for English Language Learners*, Edinburgh: Edinburgh University Press.

Quan, Z. (2016) ‘Introducing “Mobile DDL (Data-Driven Learning)” for Vocabulary Learning: An Experiment for Academic English’, *Journal of Computers in Education* 3(3): 273-87.

Scott, M. (2020) *WordSmith Tools version 8*, Stroud: Lexical Analysis Software.

Seidlhofer, B. (2002) ‘Pedagogy and Local Learner Corpora: Working with Learning-Driven Data’, in S. Granger, J. Hung and S. Petch-Tyson (eds) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, Amsterdam: John Benjamins, pp. 213-34.

Sripicharn, P. (2004) ‘Examining Native Speakers’ and Learners’ Investigation of the Same Concordance Data and its Implications for Classroom Concordancing with ELF

Learners', in G. Aston, S. Bernardini and D. Stewart (eds) *Corpora and Language Learners*, Amsterdam: John Benjamins, pp. 233-45.

Sripicharn, P. (2010) 'How Can we Prepare Learners for Using Language Corpora?', in A. O'Keefe and M. McCarthy (eds) *The Routledge Handbook of Corpus Linguistics*, London: Routledge, pp. 371-84.

Sun, Y.-C. (2003) 'Learning Process, Strategies and Web-Based Concordancers: A Case Study', *British Journal of Educational Technology* 34(5): 601-13.

Tribble, C. (1997) 'Improvising Corpora for ELT: Quick-and-Dirty Ways of Developing Corpora for Language Teaching', in B. Lewandowska-Tomaszczyk and P. J. Melia (eds) *PALC '97: Practical Applications in Language Corpora*, Lodz: Lodz University Press, pp. 106-17.

Vyatkina, N. (2016) 'Data-Driven Learning of Collocations: Learner Performance, Proficiency, and Perceptions', *Language Learning and Technology* 20(3): 159-79.

Vyatkina, N. (2020) 'Corpora as Open Educational Resources for Language Teaching', *Foreign Language Annals* 53(2): 359-70.

Wicher, O. (2020) 'Data-Driven Learning in the Secondary Classroom: A Critical Evaluation from the Perspective of Foreign Language Didactics', in P. Crosthwaite (ed.)

Data-Driven Learning for the Next Generation: Corpora and DDL for Pre-Tertiary Learners, London: Routledge, pp. 31-46.

Widdowson, H. G. (2000) 'On the Limitations of Linguistics Applied', *Applied Linguistics* 21(1): 3-25.

Widdowson, H. G. (2003) *Defining Issues in English Language Teaching*, Oxford: Oxford University Press.

Willis, D. (2003) *Rules, Patterns and Words. Grammar and Lexis in English Language Teaching*, Cambridge: Cambridge University Press.

Yao, G. (2019) 'Vocabulary Learning through Data-Driven Learning in the Context of Spanish as a Foreign Language', *Research in Corpus Linguistics* 7: 18-46.

Yoon, H. and Jo, H. (2014) 'Direct and Indirect Access to Corpora: An Exploratory Case Study Comparing Students' Error Correction and Learning Strategy Use in L2 Writing', *Language Learning and Technology* 18(1): 96-117.