# Data-driven learning: on paper, in practice.

Alex Boulton

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

ALEX BOULTON

# Data-Driven Learning: On Paper, In Practice

## 1. The impact of corpora in language teaching and learning



Corpora have much to contribute to teaching and learning, most obviously in advancing our knowledge of language and how it works, with improved descriptions finding their way into various types of reference materials. In paper form, they have been used for several centuries in preparing dictionaries, receiving considerable impetus from the COBUILD projects starting in the 1980s, with bilingual dictionaries now starting to catch up (see Cobb 2003). They are also increasingly used in the preparation of general usage manuals and specialised reference works treating particular areas of language use (such as phrasal verbs in English), as well as for grammars aiming either at comprehensive language description or at a pedagogically useful version for language learners. From such reference works with their improved linguistic description we can also expect more appropriate syllabuses firmly rooted in the reality of language use. This is most evident in the long history of corpus-based word lists, but most new materials from major publishers today claim to be corpus-based to some extent, as do more and more internationally-recognised language tests.

So pervasive is the uptake of corpus information at such levels that it is barely possible to scratch the surface, and it is likely to continue unabated. However, it is worth noting that the corpus input described so far reflects an "indirect approach" (Römer 2006: 125), in that it occurs far upstream at the level of institutions, publishers, editors, materials writers, researchers and other specialists. By the time it filters downstream to the classroom, the corpus input may have become virtually invisible to the learner. One might then wonder

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

whether corpora have a more direct contribution to make – not just *what* to learn, but *how* to learn it (Johns 2002: 110). The last twenty years in particular have seen increasing interest in the possibilities of getting learners to interact directly with corpora, especially in what Johns (e.g. 1991a) has called "data-driven learning" or DDL. This takes us far from the traditional study of an individual text or the presentation of a grammar point, and may involve significant innovation in the processes and methodologies involved.

Language teachers and learners today can access many corpora free on line. These include very large general corpora, as well as genre-specific ones for academic English and other specialisations, along with parallel corpora, comparable corpora, and learner corpora. Various interfaces also allow the user to treat the entire web as a corpus in its own right; where nothing appropriate is available, software exists to help with creating corpora from scratch (especially from the Internet), and other software can be downloaded free for corpus interrogation. Recent years have also seen the development of any number of on-line programs and software which integrate some kind of corpus consultation. The resources available are too numerous to mention, and well beyond the scope of this article.

Teachers coming to corpora for the first time may reasonably seek some kind of guidance. The closest thing to a standard "manual" devoted to DDL is probably *Concordances in the Classroom* (Tribble/Jones 1997), although there are also a number of fee-paying courses, on-line tutorials, "how-to" introductions, more general textbooks on (applied) corpus use, research papers and conferences relating direct experience, and pages of links between them; Boulton (2009a) discusses some of these. The key for many, however, is probably to experiment; this is after all the spirit of DDL itself (O'Keeffe/Farr 2003), and experience suggests that most practitioners are largely self-taught.

Given the wealth of resources available, there would seem to be every reason to rejoice as we enjoy the prospect of ever-increasing corpus use in the classroom. However, it has become commonplace within DDL circles to lament that the "trickle down" from research to teaching has not become the "torrent" predicted by Leech (1997: 2)

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

over ten years ago. Despite the considerable research interest and the multiplicity of resources available, public awareness is low: corpus consultation remains rare even in university and research environments (Thompson 2006), and it has had virtually no impact on "ordinary" learning practices elsewhere. A number of factors may account for this.

Firstly, it may be that DDL itself does not live up to its claims if it is found to be too difficult, demotivating, irrelevant, and inefficient. Boulton (2009a) discusses these and other barriers in some detail; suffice to say here that current research is, on the whole, positive, with participating learners enjoying the work and benefiting from it (cf. Boulton 2008a). Those who voice these objections are typically working teachers, suggesting a lack of communication between the research and teaching communities, as well as deeper concerns such as the perceived threat to the teacher's role, especially a loss of power, control, and respect as the ultimate knower. Another set of objections concerns the resources themselves: the corpora and software are not always appropriate for learning purposes (cf. Kosem 2008), often with "too many degrees of freedom […] for the ordinary learner" (Schmied 2006: 104). Computer rooms may be unavailable when needed, badly-equipped, too small, subject to breakdown, lacking in technical backup, or simply non-existent. These problems are certainly very real in many cases; but again, the suspicion is that they reflect a deeper underlying malaise on the part of the teacher, especially resentment of new technology and the time spent mastering it, as well as the risk to face in front of learners who are possibly more literate than the teachers in ICT (information and communication technology). This is a teacher's version of the student's "technophobia" cited by Seidlhofer (2000: 208) and others, and such teachers are likely to be hostile to any use of ICT or CALL (computer-assisted language learning).

How then are we to counter such objections and promote DDL in wider circles? Römer's (2006) "wish list" includes more relevant corpora and more user-friendly software designed with language learning in mind, as well as her "corpus mission" in the form of better communication between current practitioners, between researchers and teachers, and especially the integration of corpus consultation into

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

teacher-training courses. On the last point, most initiatives to date involve in-service programmes or MA courses where students may go on to train as teachers. But if Conrad (2000: 556) is right in that "the strongest force for change could be a new generation of ESL teachers", then DDL needs to be incorporated into pre-service training. This is rare at the moment for understandable reasons (though see Farr 2008; O'Keeffe/Farr 2003; Seidlhofer 2000): student teachers are likely to be more interested in the requirements essential to qualifying. As long as DDL is absent at this level, it is likely to be seen as marginal, or even as an unnecessary extra burden (cf. Mauranen 2004: 197). A similar comment can be made in relation to classroom use: many learners are most concerned with passing their exams and gaining qualifications, so may perceive DDL as not directly relevant or even a waste of time.

To sum up so far: for DDL to make any significant impact, it has to be introduced early, reduce perceived threats to teacher (and learner) roles, circumvent the problems inherent in using computers, and enhance its reputation for direct relevance and ease of use. Which poses something of a paradox: on the one hand, DDL is *not* ordinary practice; on the other, DDL has to *become* ordinary practice (or at least, not to be seen as *extra*ordinary when first encountered). There might then be an argument, contrary to common belief, for presenting DDL not as a radical new technique, but as ordinary practice alongside other ordinary activities and materials – in other words, to "demystify" corpus use (Gabrielatos 2005). The integration itself should make things easier for learners and teachers by forcing us as researchers to find ways to reduce some of the more radical aspects, to eliminate excess baggage rather than continually seeking to add new features, and should also by association help to identify DDL with ordinary practice. All of this needs to be done without losing the advantages of DDL, and without it becoming completely invisible or melting into the background.

One direction lies in the development of ICT and CALL resources which integrate corpus consultation in some form or another, and countless such applications can be found on the Internet. These initiatives are hugely appreciated, but one suspects they are also

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

underused outside the environments where they were developed. Computers have enormous appeal for some, but may deter many others: users may not be well-disposed to any use of new technology, and resent spending time finding out how to navigate each piece of software or Internet interface. ICT has enabled increased interactivity, undoubtedly a worthwhile objective, but many teachers want to be able to print out activities for classroom use – a problem even with some static html interfaces. In other words, ICT applications are perhaps not ordinary enough.

In the classroom itself, the most ordinary materials besides pens and paper are probably coursebooks. One obvious possibility then would be to integrate elements of DDL into such a medium. This would inevitably entail some watering down of the hard-core, hands-on, autonomous approach to consultation of electronic corpora; but if this can be achieved while retaining at least some of the benefits of DDL, the compromise may be worthwhile.

## 2. DDL and the print medium

A major question then is whether coursebooks and other "off-the-peg" resources can successfully integrate a DDL approach – or indeed whether they can truly constitute DDL at all (e.g. Bernardini 2001: 228). The second element may seem intractable in the absence of any watertight definition of DDL, although it might be pointed out that if Johns (e.g. 1991a, 1991b), widely considered as the father of DDL, made extensive use of printed handouts, then it is difficult to maintain that they are not DDL. If appeals to authority are considered of dubious legitimacy, one might also mention that providing learners with printed data and accompanying activities is probably "the most common procedure" (Todd 2001: 93) for corpus use in language teaching. Dozens if not hundreds of research papers report such practices, while the hands-on activities in countless others could just

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

as easily be done on paper (e.g. the links to precast concordances used in Gaskell/Cobb 2004).

As Johns (1988: 14) pointed out, DDL "entails a shift in the traditional division of roles between student and teacher, with […] the teacher acting as research director and research collaborator rather than transmitter of knowledge." This key element holds true for prefabricated materials, even though the answers may be known in advance – the "rule-hiding" Johns (1991a: 4) acknowledges in his own handouts. Perhaps they do not, as Johns (1991b: 30) put it, entirely "cut out the middleman" (i.e. the teacher), but the teacher takes on the new role of guide, and is certainly not the "bad old magister/tutor" feared by Cobb (1997: ch2). Printed materials have the substantial advantage of obviating the need for computer laboratories and the associated problems mentioned at the start of this article. More positively, they may actually improve the efficiency of the process by reducing some of the difficulties associated with hands-on work, especially the risk of being "overwhelmed" by the mass of data (Johns 1986: 156), much of it irrelevant, incomprehensible, and extremely messy. With prepared materials, the data can be sorted and grouped appropriately, and carefully devised activities can eliminate much of the tedium associated with hands-on work, and rule out irrelevant paths from the start. Learners may react more favourably to this compromise: Granath (1998) found that less than half of her students liked deciding the queries themselves, while two thirds appreciated the teacher-designed exercises; and Whistle's (1999: 77) students simply did not see why the concordances could not be printed out in advance. After a brief introduction to paper-based concordances, 40% of the students in Boulton's (forthcoming) study expressed no opinion as to whether they would like to try it on computer, while the others were evenly split for and against.

If printed materials have these advantages, then it makes little objective difference whether they are provided by teachers or by coursebook writers. Of course, the printed materials referred to in scholarly papers are largely "reactive" in the sense that they are created in response to specific questions and problems among a given population of learners. However, they are often recycled with other

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

learners (and even by other teachers), given the obvious reluctance to abandon them after the considerable time spent in their preparation (cf. Warren 1998: 214). Use of published materials only increases the distance slightly, and as far back as 1984 Johns was suggesting concordances could be integrated directly into teaching materials (Higgins/Johns 1984: 93), later revealing that his own "experience in using concordance data reactively has indicated that it could be used proactively also in a more traditional teacher-centred setting" (Johns 1991b: 31). From the learners' point of view, a crucial element of DDL is to be able to take greater responsibility for their own learning; providing handouts clearly reduces the scope for this. However, the basic process still consists of exploring the data, detecting patterns, formulating hypotheses and generalising to other cases. In other words, learners do still have more input than in traditional teaching, and the compromise may be more appealing to those who do not have a particularly inductive style to start with.

Materials exist to make the teacher's job easier and more effective, removing some of the burden in terms of time, effort and know-how, in addition to providing the resources themselves. In turn, the teacher's role consists, in part, of making the learners' task easier in much the same ways; they will be failing in this if DDL is perceived as making things unnecessarily difficult, which may be the case where learners are introduced all at once to the new approach (DDL), new materials (corpora), and new technology (software). DDL as an approach can seem difficult enough, with its associated elements of discovery learning and induction, not to mention the fuzzy and probabilistic nature of language – quite different from the familiar comfort of rules and "being taught". The use of corpora brings additional problems due to issues of authenticity, decontextualisation, (ir)relevance, quantity, truncation, and so on – a far cry from reading or listening to a text, invented or not. Additionally, the new technology represents a formidable barrier as learners negotiate technical aspects of the interface, formulate and refine workable queries in the appropriate query syntax, even assuming they have access to computers that can handle the tasks and that technical support is available.

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

Kaltenböck and Mehlmauer-Larcher (2005: 81) argue that the lack of mediation is a major reason for the failure of DDL to make greater impact, claiming it as a "necessary prerequisite for successful application of computer corpora in language teaching." If it is possible to simplify the equation in initial stages, so much the better. Of the three elements – the DDL approach, the corpora, the technology – the last is probably the most commonly-cited source of difficulty (e.g. Farr 2008), so would seem to be the one best left for later; the use of printed materials allows precisely this (cf. Lamy/Klarskov Mortensen 2007: §4.1). Chambers and Kelly (2004: 128) remind us of the "truism that technology is at its most successful when the technology disappears", also citing the over-reliance on technology as one of the factors inhibiting the spread of DDL to a wider public.

All of this is not to suggest hostility to hands-on corpus consultation: quite the opposite. Using technology may be more environmentally friendly than paper materials, allow far greater autonomy, and be motivating for many – although perhaps fewer than is generally assumed: Jarvis (2004) found only 8% of respondents in higher education in Britain definitely agreeing that computers were motivating. However simple the corpus interface, however well DDL is integrated with other functions, however user-friendly the program – the very fact of having to use computers will deter many. Perhaps the main argument for introducing DDL via printed materials is that it cuts out this barrier and thus has the potential to open it up to a wider audience. This in turn will hopefully set the scene for later work on computers as learners gain the knowledge and skills necessary for more autonomous work – choosing the corpus and software, deciding the language points to work on, adapting the approach to their individual needs, styles and preferences. If autonomy has often been singled out as the main advantage of DDL (e.g. Aston 2001: 41), it should be remembered that autonomisation itself is a gradual process (cf. Mukherjee 2006). "Autonomy can still be engendered where concordances are provided as materials by teachers […] DDL can still promote learner autonomy even in a less than ideal environment" (Allan 2006: 15). Using printed materials allows learners to take things at their own pace, one step at a time (Turnbull/Burston 1998:

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

12), with correspondingly less chance of being put off by excessive demands. A gradual initiation should allow them to develop the necessary techniques before going on to find or create their own corpora and locate appropriate software on the Internet, at which stage they will be able to continue their language learning outside the classroom and after their education has finished without the need for teachers or textbooks. Initial use of prepared materials does not imply that hands-on concordancing, with everything that entails, remains the long-term objective for those who continue to need foreign language skills.

Theoretical arguments aside, the crucial issue is whether there is benefit to the learner. Although many scholarly articles discuss uses of printed materials, there is surprisingly little concrete research: a survey of empirical DDL studies (Boulton 2008a) found only a handful analysing learning outcomes from use of printed materials. Ciesielska-Ciupek (2001) is unusual in that she was also working in a secondary-school environment, although the experiment design and data analysis do not allow more than a subjective appreciation of the positive outcomes. A rigorously statistical large-scale study by Koosha and Jafarpour (2006) found the DDL group making substantially greater gains in the target items than the control group. Allan (2006) similarly gives the advantage to DDL, although she was working with far fewer students and also admits certain design problems. Intriguingly, she provides some evidence that her learners also performed better on non-target items, suggesting that the *process* of DDL leads to greater language awareness, noticing skills, and ultimately better learning – even from paper-based resources. Finally, three controlled experiments by Boulton (2008b, 2009b, forthcoming) show learners making significant gains on target items in post-tests, although differences with control groups were mostly small or not significant. However, the learners in these studies are experiencing their first taste of DDL with no prior training, so the results not only show that DDL can lead to immediate learning on a par with traditional approaches, but also suggests that training and further experience would give it a distinct advantage over traditional teaching, even at lower levels. Although no studies to date directly compare the

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

benefits of hands-on corpus consultation with those of prepared materials (cf. Chambers 2005: 121), it does seem that DDL can be useful via the printed medium.

## 3. Existing DDL materials in print

This section looks at a number of printed course materials for English as a foreign language (EFL). This would seem to be the most likely place to find some aspects of DDL, though it is probable that there are at least some additional items available for other languages, countries and educational environments (see e.g. Tono 2008). Nonetheless, work in and on English is likely to be predominant, as major DDL events and research tend to be conducted in English, which is also the target language most frequently discussed there. Furthermore, there is a good chance that an adult market would be strongly represented as DDL is still largely associated with higher education.

The aim is not to provide a comprehensive review, but to focus only on those elements relevant to DDL. Rather than attempt a rigorous definition of DDL which might miss some interesting items, it seems preferable to cast the net fairly wide. All of the materials here claim to be using *authentic* data obtained from *corpora*, and these data are the source of *learning* – in other words, some kind of inductive approach to the corpus data is required. These elements are not sufficient for activities to be called DDL (cf. Gabrielatos 2005), nor are they strictly necessary (there may be DDL-type activities where they are absent); but they seem to be at the heart of what is generally accepted as DDL (Boulton 2009c).



### *3.1. Collins COBUILD English Course 2. Willis/Willis 1988.*

The *Collins COBUILD English Course* was based on a lexical syllabus, a list of words and phrases and their uses derived from the

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

COBUILD corpus, then a fraction of its current size. The concept of the lexical syllabus was proposed by Sinclair and Renouf (1988), outlined further in Willis (1990), and extended by Lewis (e.g. 1993, 1997). It does not ignore grammar and other areas of language, but they are subordinate to the central organising feature of lexis – "the commonest word forms in the language; their central patterns of usage; the combinations which they typically form" (Sinclair/Renouf 1988: 148). The corpus input provides some justification for the prominent phrase on all COBUILD products at the time, "Helping learners with *real* English." A corpus-informed lexicon features at the back of each edition: Level 2 adds 850 new words to the 700 covered in level 1, extensively treated and recycled throughout; level 3 aims for 2500 words in total. The authors aim for a task-based approach, and the back cover proclaims that learners will "discover recurring features of the language by analysing samples of real English." This is at the level of individual texts taken from the corpus or elsewhere, and while learners are frequently required to match extracts from the texts or dialogues against the grammar rules provided, this can be argued not to involve induction as such. Other activities require learners to categorise words, phrases or sentences, or to identify common factors, which certainly encourage noticing skills and language awareness. While some of these fragments are taken from the corpus, no use is made of concordances.

The COBUILD course sold reasonably well without being a runaway success. The authors (cited in Schmitt/McCarthy 1997: 323) attribute this largely to "packaging", essential for any innovation to reach its public. For example, grammar is treated implicitly as a consequence of the main lexical syllabus; teachers and learners generally expect a stronger grammar profile. They also feel a more eclectic approach would have been useful to "enable innovation to take place within a relatively familiar environment". Although new editions were never produced and it is now out of print, the series proved influential beyond its sales. It is doubtful, however, whether it can really be regarded as DDL.

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

*3.2. Touchstone 4. McCarthy/McCarten/Sandiford 2006.*

A more recent series of coursebooks and one of the new generation of "corpus-informed" materials is *Touchstone*, based on the Cambridge International Corpus and designed especially for the American market. The corpus element is given prominence on the back cover and in the introduction, as well as in the associated publicity and accompanying website, and in a monograph about the book (McCarthy 2004). The corpus input helps in deciding what to teach at different stages, and provides a source of texts for the course. However, access to the corpus is at all stages "mediated" by the writers, who may adapt or modify texts as they see fit, "building" texts and "constructing" dialogues out of the original data; they are thus (only) a "reflection of real usage" (McCarthy 2004). With the exception of a list of the 500 most frequent words, the corpus input is largely invisible in the book itself, part of a deliberate decision to produce materials that "are familiar in structure and easy to use" (McCarthy 2004: 15).

The introduction claims the course is based on "communicative methodologies"; most relevant of the six main features for the present article is that "it promotes active and inductive learning" (p. vi). The most evident example of this is in the "figure it out" sections at the beginning of each unit, where learners are encouraged to focus on the grammar point and work out the meaning and use for themselves prior to reading the explanation on the facing page. The texts themselves are clearly carefully designed to present the grammar point as saliently as possible, and to contrast it with known structures. Similar points can be made about vocabulary which, unusually for a coursebook, is given particular prominence here (McCarten, 2007, provides the rationale for this).

Although the course is corpus-informed, the language presented bears little trace of its corpus origin; the texts are so mediated that, even where induction is called for, the answers are absolutely transparent. These are deliberate choices, and the authors make no claims to a DDL approach – indeed, quite the opposite: "Teachers and learners should expect that, in most ways, corpus informed materials

will look like traditionally prepared materials" (McCarthy 2004: 15). In other words, absence of DDL can in no way be taken as a criticism. The importance of corpus-informed materials such as *Touchstone* can scarcely be overstated; the authors have produced a major step forward, and the publishers have clearly put enormous resources behind it; but again, it is not DDL.

### *3.3. The Intermediate Choice. Mohamed/Acklam 1995.*

Corpora receive greater prominence in this course: unit 1 features activities based on an interview with a researcher talking about the British National Corpus (BNC) and its uses in language learning. This sets the scene for 18 short concordance extracts throughout the book (average 6½ lines); these are sometimes called "sentences", though they are invariably in the KWIC (keyword in context) format, with the keyword manually highlighted and the important surrounding text in bold. These concordances may present multiple uses of the same item, but frequently feature a number of different but related items (e.g. unit 13: *a few*, *a little*, *any*, *some*, *many*, *much*); here each occurs in one or two lines only, and essentially provides a novel visual format to the traditional function of "example".

The accompanying tasks are often deductive (to categorise the concordance lines according to given rules), occasionally inductive (asking learners to come up with their own categorisations, or to answer questions based on the concordances). The teacher's book provides some rationale for this, and further emphasises the importance of spoken material from the BNC in every unit. While the overall feel is distinctly close to DDL, the corpus extracts are largely to illustrate rules, with little opportunity for the learner to really explore concordances; this is borne out by the lack of any overt corpus material in the accompanying workbook (Thornbury 1995).

### *3.4. Focus on Vocabulary. Schmitt/Schmitt 2005.*

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

*Focus on Vocabulary* has a very explicit aim, as its subtitle makes clear: "mastering the Academic Word List". The AWL was devised by Coxhead[1] (2000), inspired by the University Word List (cf. Nation 1990) which in turn supplemented the largely frequency-based General Service List produced by West (1953). The 570 words on the AWL are taken from a rigorously constituted corpus of a variety of academic texts; they account for 10% of the corpus without being domain-specific, and thus constitute a valuable supplement to general lists for learners needing English for academic purposes (Coxhead 2000: 222). *Focus on Vocabulary* aims to cover over 500 of these words systematically, an ambitious endeavour for a single book.

A number of activities require learners to guess the meanings of words, or to focus on collocations, usage patterns, and so on – the types of activities commonly associated with DDL. These activities are based on two to four full sentences, presumably considered to be more useful or less intimidating than KWIC concordances, and are apparently taken from the corpus. Other activities encourage learners to detect patterns of usage, and especially of word families. While the presentation of the data is thus not typical of DDL, it seems that the writers are aiming in this direction, and the results of other recent research are incorporated at all stages: as the back cover claims, it really is a "research-based vocabulary textbook", both in its conception and in what the learners are required to do.

### *3.5. Natural Grammar. Thornbury 2004.*

Thornbury is well aware of issues related to deduction and induction, which receive a chapter each in *How to Teach Grammar* (1999); in the latter section he includes a sample lesson using KWIC concordances to teach verbs that take either the infinitive or the *–ing* form. His *Natural Grammar*, examined here, is more than just a grammar book, as the introduction explains:

1 The AWL and the GSL can both be downloaded from <http://www.lextutor.ca/freq/lists_download/>.

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

> As Professor John Sinclair put it: 'Learners would do well to learn the common words of the language very thoroughly, because they carry the main patterns of the language.' […] By learning these high-frequency words and their high-frequency patterns, the learner is getting traditional grammar 'for free', as it were. (p. i)

*Natural Grammar* also pays homage to other researchers from the COBUILD tradition, and clearly owes a debt to the lexical syllabus concepts behind the *Collins COBUILD English Course* (see above). The book presents 100 of the commonest words of English – mostly grammar-function words – in alphabetical order, so is not a coursebook as such, although the introduction offers little advice for approaching the book. A third of the units contain an exercise where the learner is asked to work with somewhere between 9 and 19 "concordance lines", which in all cases are full sentences and not KWICs. The task in these exercises is to match each sentence against a grammar pattern already provided; in other words, they are exclusively deductive activities. Working with multiple contexts like this is clearly very much DDL-inspired, although the deductive exercises hold it back from going the whole way. This might be a deliberate attempt to avoid something too dramatically new, or perhaps is intended to reduce false inferences or potential time wasting.

### *3.6. Phrasal Verbs: American English. Barlow/Burdine 2006. Business Phrasal Verbs and Collocations. Burdine/Barlow 2008.*

Both these books bear the label *CorpusLAB*, and the word *corpus* features prominently on the front cover, while the back claims "a new approach to language learning" featuring "corpus-based instruction", among other things. Frequency information is given separately for spoken and written English for each target verb. Every unit begins with the instruction to "study these examples", exclusively full sentences with the main meanings provided. However, other exercises ask the learners to examine a short set of concordance lines, followed

by some guiding questions focusing the learners' attention on patterns of meaning, usage and collocation. Exercises such as these are described in the introduction as "pattern identification" and "concordance-based research" activities (2006: 4). The later book omits the second phrase, but replaces it with mention of "a technique called data-driven learning in which you will analyse and classify usage" (2008: 3).

In the earlier book, the data in these exercises consist entirely of full sentences, between three and eight for each activity; these are occasionally aligned around the target words in bold. The data in the second book are presented as screen shots of between 12 and 15 KWICs, with the target words in bold face. In the second book, these activities feature systematically in the review section titled "CorpusLab exercises" after every ten units, meaning that they are introduced earlier than in the first book where they are scattered among the later units. These activities are certainly DDL, but are comparatively infrequent: there are only nine in the first book and five in the second. They do however have a higher profile in the second book as they are introduced earlier and more systematically, and the KWIC presentation is visually more remarkable.

### *3.7. Exploring Academic English. Thurstun/Candlin 1997.*

*Exploring Academic English* is entirely given over to recognisably corpus-based materials and DDL techniques. This is possible in part due to its very concentrated focus on a small number of non-domain specific rhetorical vocabulary items drawn from Nation's (1990) *University Word List*. It is only concerned with academic English for essay writing at university level, and is primarily a workbook rather than a coursebook; the underlying rationale is outlined in Thurstun and Candlin (1998). A small group of rhetorically related items is briefly introduced, then each one is then given the same systematic treatment. The first "look" phase presents an entire page of about 30 KWIC concordance lines for learners to study on their own; in the second "familiarise" phase, they have to use this information to

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

answer questions about different meanings, uses, collocates, colligations, and so on. Learners play an active role here – extracting data from the KWICs, identifying patterns, grouping information appropriately – which calls upon both inductive and deductive processes. The third "practise" phase involves mainly cloze and matching exercises, included precisely because they "provide a sense of familiarity given the novel nature of the materials" (Thurstun/Candlin 1998: 273). After each group of units, multiple gapped concordances are provided along the lines of those piloted by Stevens (1991). The final "create" phase asks learners to write a short text recycling the target items.

Each of the six units focuses on only three or four target items, a relatively low return for an average of twenty pages – one of the major criticisms in Thompson's (2001) review of the book – but a deliberate choice as the authors make clear (1998). The advantage is that the target items are treated in considerable depth, with learners receiving repeated exposure to them as well as to considerable quantities of related language. As each unit is based on a rhetorical function, learners are likely to become more sensitive to other uses, and indeed to improve their noticing and thus learning skills in general (cf. Allan 2006). Finally, the book's systematic approach and tips in dealing with complex data can be treated as an introduction to DDL in itself, "train[ing] the learner in effective corpus analysis skills" (Thompson 2001: 30) and thus facilitating a transition to hands-on DDL. Thurstun and Candlin (1998: 277) also report that students piloting the materials "overwhelmingly indicated that they find all exercises 'very helpful' or 'helpful'."



*3.8. Concordance Samplers 2: Phrasal Verbs (CS2). Goodale 1995a.*

COBUILD materials gained an early reputation for their uncompromising rethink of the language and of language learning based on evidence rather than pre-existing ideas; the *Concordance Samplers* are as innovative as one might expect. In addition to the *Phrasal Verbs* volume discussed here, two others were produced on

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

*Tenses* (Capel 1993) and *Prepositions* (Goodale 1995b)[2]. An introductory needs test is based on multiple concordance lines just like Stevens (1991) and also found in *Exploring Academic English* (see above); in a final series of revision tests, each line is a separate test item. Following a very brief "guide to meanings" for each particle, the main body of the book is taken up with 46 pages of KWICs at 40 lines per page: the 16 particles have up to seven pages each, while 10 major verbs each have a page to themselves; the concordances are usually but not always left-aligned. Although the data are presumably selected rather than representing a random sample, apart from that they are "completely unedited", as Sinclair points out in the introduction, since otherwise their "freshness and […] authenticity will diminish" (p. 4).

The final pages feature five worksheets which are not specific to any particular phrasal verb – some might even be adapted to other language points. This does mean that the learner has some quite mechanical tasks to perform, for example listing all the prepositions that follow a given phrasal verb, or all its separable occurrences or passive forms, or grouping different meanings, etc. While one might wonder whether such an approach would be too tedious and laborious on paper, Hadley (2002) used the book with low-level Japanese learners and reports finding their interest and motivation increased. The overall feel of the *Concordance Samplers* is closer to hands-on DDL: more data, fewer exercises, less mediation, with more of the responsibility falling on the learner (who may as a result learn more and become more autonomous).

### *3.9. Alternative sources*

Published courses are not the sole repository of printed DDL materials. Research publications constitute one possibly underestimated source, as many describe particular courses and

2 There is some variation between the different editions; for example Capel (1993) provides the worksheets at the start of the book, and has only one set of tests resembling the revision tests at the end of the book discussed here.

include examples of worksheets used. These sources are too numerous to name individually, but many can be found in academic journals and in collections of papers such as those issuing from the *Teaching and Language Corpora* (TaLC) conferences. Similarly, a number of items aimed at teachers contain examples of printed materials; particularly notable here is Tribble and Jones (1997), but useable activities can be found in several other "how-to" introductions, courses, tutorials and books (see Boulton 2009a). Textbooks on corpus linguistics can also provide inspiration, even if they are not primarily aimed at language learners; an especially valuable resource in this respect is Sinclair's *Reading Concordances* (2003).

Even though some of these materials are in printed format ready for use, most are unlikely to reach teachers directly. This has led some researchers and teachers to post their materials on line, Johns again setting the trend. Worksheets in his *Virtual DDL Library*[3] focus on wide-ranging aspects of lexis, grammar, meaning, usage, discourse, and so on. Most involve multiple concordances, usually in the KWIC format but sometimes complete sentences, some even from parallel corpora. Occasionally the learner is given a traditional description or explanation in advance (such as might be found in any dictionary, usage manual or grammar book); the task is then to test this description, or to categorise concordances according to the description. More usually, the data are accompanied by guidelines to help the learner focus on the target item, ask relevant questions, detect the patterns of use in the contexts, and formulate appropriate inferences. Finally, activities are provided for learners to apply what they have found; what is remarkable here is the variety of activities, many of which are traditional but here based on authentic data in the form of individual or multiple concordances: identifying and underlining target items; cloze and other forms of completion exercises; choosing the right form in context; putting bare items in the

3 Since 2007, this has been hosted at <http://www.eisu2.bham.ac.uk/johnstf/ddl_lib.htm>. It also includes sample activities by Joseph Rézeau, although these dead links have to be traced via an archive such as <http://www.archive.org>.

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

appropriate form (e.g. tense, aspect, countability); correcting inappropriate forms; matching split sentences; re-arranging items; word-formation (affixation, compounding, etc.); question/answer (e.g. *what's the difference between X and Y?* or *what do X and Y have in common?*); grouping lines according to meaning, usage, etc.; translation (especially but not exclusively in the case of parallel concordances); writing sentences or inventing new examples; and so on. Overall, most of the materials here conform to the discovery learning or inductive paradigm of (i) observation, (ii) hypothesis-formation, (iii) use/experimentation; one of the most developed versions of this can be found in Willis (2003).

A somewhat different format also proposed by Johns (2002)[4] is the "kibbitzer", the analogy being to chess with onlookers providing comments from a distance. The starting point is students' academic writing, and the comments are based on learners' questions, teachers' corrections, or Johns' own reactions. Rather than simply telling students the answers (even where this is possible), the idea is to lead learners through the stages of querying a corpus to find answers together (cf. Johns 1997). Kibbitzers tend to be based on very specific points, but this site contains notes based on 77 real examples from 1996 to 2000, many of which cover quite common points and so can be reused or adapted. The idea continues with various kibbitzers written by Swales and colleagues at MICASE[5]. The 14 examples here are generally far longer and more complete than Johns' notes, and notably include data other than just concordances, especially in the form of frequency information, collocates tables, and graphs comparing distributions across genres, between sexes and age groups, between corpora, and so on.

A number of other individuals or groups have put printable materials on line, of which the following are just a few. Estling



4 See also *Tim Johns' EAP page*: <http://www.eisu2.bham.ac.uk/johnstf/timeap3.htm>.

5 J. Swales, A. Ohlrogge, A. Adel, F. Reinhard, J. Kruis, J. McCormick, J. Tsang, R. Alejo, R. Maybaum, S. Pilon, S. Richardson, S. Shryl Leicher & S. Marx. *MICASE kibbitzers* <http://lw.lsa.umich.edu/eli/micase/kibbitzer.htm>.

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

Vannestål and colleagues[6] provide slides and exercise booklets for an introduction to corpus use for grammar classes (described in Estling Vannestål/Lindquist 2007). Sripicharn has over 20 exercises on individual points, all taken from the freely available sampler of the Bank of English[7]; these can be printed or done very simply on line. Materials by Lopes Moreira Filho[8] also include a number of basic interactive exercises, all with instructions in Spanish. Barlow has recently set aside part of his *CorpusLAB* site[9] for teachers to upload their own DDL materials, and it is to be hoped that more and more resources will become available there. Chambers and Kelly (2004) report several planned projects to develop online worksheets, and a number of other individual sites contain printable and reusable materials.

## 4. Discussion

Not all of the materials discussed above might be considered DDL: it is not enough to be corpus-informed, or to include extracts taken from a corpus, or to use inductive learning. In the case of published materials, the use of corpora tends to be given a high profile in the accompanying publicity, on the book covers and in the introductions, but this visibility is often lost in the materials themselves where the extracts take on the familiar form of complete texts or sentences, even



6 M. Estling Vannestål, H. Lindquist, E. Tyberg, S. Månsson & M. Karlsson. *Corpora in grammar teaching* <http://www.vxu.se/hum/utb/amnen/engelska/kig/>.

7 P. Sripicharn, *My DDL Materials* <http://www.geocities.com/tonypgnews/units_index_pilot.htm>. *Collins WordbanksOnline English corpus* <http://www.collins.co.uk/Corpus/CorpusSearch.aspx>.

8 J. Lopes Moreira Filho. *Reading class builder* <http://www.corpuslg.org/software/rcb/materiais.html>.

9 M. Barlow. *CorpusLAB* <http://www.corpuslab.com/>.

if these are called "concordances" in some cases. The choice of terminology is presumably a deliberate one: "sentence" would seem to keep things familiar and reduce jargon, while "concordance" plays on the novelty of the activity and enhances standing among the research community.

If the net was initially cast wide, it is partly because there are so few promising items: e-mails sent to major EFL publishers did not bring up any further materials, nor did postings to CorporaList and CorpusCALL[10]. Furthermore, several of the items discussed here are old or out of print, while others have limited distribution from small publishers. The fact that most have merited reviews and are the subject of research articles by the authors and others suggests that they are both rare and innovative, even ground-breaking.

An essential question arises: why does so little published material make use of DDL? The immediate answer has to be commercial: language teaching materials represent "big business" (Cook/Seidlhofer 1995: 8), and publishers need to be convinced that such materials will sell well. It is important to underline that this has little to do with any pedagogical merit of DDL: it is simply difficult to blame publishers for being reluctant to risk investing in materials if, having done their market research, they find the market does not exist.[11] If past experience is a factor, we must assume that the materials described here have not enjoyed the commercial success needed to inspire new investment from major publishers. In most cases, this is not surprising, as they are from minor sources with limited publicity, and are often aimed at very specific and hence small segments of the market. Of the two general coursebooks from major publishers, the *Collins COBUILD English Course* did sell in large



10 Several responses did bring to light a number of on-line interactive resources which there is not the space to discuss here; summaries can be found at <http://www.uib.no/mailman/public/corpora/2008-April/006422.html> and <http://www.jiscmail.ac.uk/cgi-bin/webadmin?A1=ind0810&L=corpuscall&X=157DDC7ED291618E37&Y=boulton%40UNIV-NANCY2.FR>.

11 However, responses to recent email enquiries suggest that representatives of major publishers are often quite unaware of what DDL is.

numbers, and *Touchstone* seems set to be a major commercial success; but tellingly, neither really promotes a DDL approach.

Hanks (2008: 221) forcefully makes a similar point regarding dictionaries, but in the following quotation the word *dictionary* might easily be replaced by *language coursebooks*:

> Dictionary publishing is characteristically caught in a vicious circle. It is a cut-throat competitive business, in which marketing is at least as important as content. […] Dictionary publishers tend to pride themselves on being 'market-driven'. This is the root of a problem. Existing dictionaries create certain expectations among users about what dictionaries will be like. These expectations are conservative; people expect new dictionaries to be improved versions of old ones, not radical new departures. How could it be otherwise? […] So dictionary publishers are typically conservative, driven by an unthinking market and opposed to any innovation that might frighten away buyers.

There is clearly a catch-22 situation here which applies equally to DDL: materials are needed to create a market, but without an existing market publishers are reluctant to take the risk. However, studies of attitudes among "key players" (textbook writers, teachers, teacher trainees, and teacher trainers), such as that conducted by Heyvaert and Laffut (2008) in Flanders, suggest changes may not be far off.

One problem is that DDL practitioners tend to be primarily concentrated in higher education rather than within the larger markets of secondary education or language schools. One may note that many of the publications discussed here are the work of researchers or of teachers intimately connected with a research environment. This is no doubt inevitable, insofar as new ideas tend to be taken up first in a research environment where practitioners are expected to combine teaching and research interests. Creating new software and experimenting with new techniques is not only part of the job, it is for many the most interesting aspect of the job. It is part of what attracts people to university work in the first place, and there is considerable pressure to publish research for career purposes (textbooks often do not "count" in assessment exercises). This means that researchers may be more interested in doing new things rather than consolidating

current work by spreading existing ideas to a wider audience. The impetus for innovation is reinforced by the specialised contexts of higher education, which means that there are often specific needs or circumstances not catered for in existing materials; indeed, extended use of published materials might even be disparaged in research environments. There are also fewer outside constraints to hinder innovation, with syllabuses and course contents decided by the individual or at a local level. Finally, there are the resources to make it possible, in terms of know-how, hardware and software, class sizes, and perhaps most importantly, time available.

As a consequence of all this, most of the techniques, activities, corpora, software and so on have been designed with the university environment in mind, with comparatively little energy devoted to adapting the approach to other contexts. This reinforces the idea that DDL is only appropriate for adult, sophisticated, advanced university-level learners, although what little research there is with other types of learner tends to be largely positive (cf. Boulton 2008a). The situation is unlikely to change as long as DDL remains the domain of university teachers with their strong interest in research. Input from full-time working teachers is essential (cf. McCarthy 2008), but they cannot reasonably be expected to make the crossover themselves: the onus is on the researchers to build the necessary bridges – partly through providing more accessible materials.



## 5. Perspectives

One of the most immediate solutions is to continue sharing resources, especially via the Internet. They can be difficult to find if scattered around the web on individual homepages, and it may be useful for each to link to other sites, or to have centralised pages of links[12].

12 Some existing sites include Tom Cobb: *Compleat Lexical Tutor* <http://www.lextutor.ca/>; David Lee: *Devoted to Corpora*

Alternatively, it might be better to group them directly on high-profile sites such as *CorpusLAB*; comparable initiatives exist for other corpus resources, such as the *Oxford Text Archive* or the *Common Language Resources and Technology Infrastructure Network*[13]. A further advantage of a centralised resource is that the qualities and failings of different materials become more apparent, which can provide inspiration for improving existing resources, creating new ones, and filling in the gaps.

This DIY approach is to be lauded, but there are actually fewer worksheets ready to be printed out for immediate use than is sometimes claimed (e.g. O'Keeffe et al. 2007: 24), and in any case they can only take us so far. The language points covered are extremely heterogeneous, as they tend to be based on particular points of difficulty which have struck individual teacher-researchers, with little connection between them. Similarly, variety may be a good thing in itself, but the huge diversity of different types of instructions and activities can appear confusing. In any case, the goodwill of individuals has its limits. Creating materials can be extremely time-consuming even for one's own use: Johns (1991a: 4) spent four hours preparing a handout for class use, and as long again to make it presentable for inclusion as an appendix to a research article. Where the aim is to share, instructions need to be completely transparent and generalisable to other contexts, and potential contributors may be concerned to produce perfectly formatted worksheets if they feel they may be judged on the result. Finally, some people may be unwilling to share the fruits of their considerable labours for free, guarding their materials jealously; this is also a problem in corpus creation, where copyright is a further issue, and one which has yet to be resolved definitively even for the extracts used in not-for-profit resources.

More decisive in promoting public awareness is what publishers do. Coursebooks entirely devoted to DDL present a number of



<http://devoted.to/corpora>; Michael Barlow: *Text Corpora and Text Linguistics* <http://www.athel.com/corpus.html>; Betsy Kerr's "useful links" <http://www.tc.umn.edu/~bjkerr/CSC_DDL_Bib.htm>.

13 OTA <http://ota.ahds.ac.uk/>; CLARIN <http://www.clarin.eu/>.

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

problems, not least that they ignore alternative learning styles, and that an overdose of DDL can be demotivating if it becomes too repetitive and mechanical, as has been remarked elsewhere (e.g. Thurstun/Candlin 1998). Variety is important pedagogically speaking, and purely DDL resources might be best kept in reserve as supplementary materials for specific language points; *Exploring Academic English* and the *Corpus Samplers* discussed above may find their best use this way. A further possibility would be to focus on language items where learners are known to have difficulty. These might be identified from learner corpora, and photocopiable worksheets developed for individual use as appropriate. In other words, where the deductive approach of traditional teaching is found wanting, an alternative inductive DDL approach might have something to contribute (cf. Boulton forthcoming).

On the whole, it will probably be more fruitful to find ways to integrate DDL activities into coursebooks, workbooks and photocopiable supplements, with tips for use and extra activities included in teacher's books (cf. Chambers/Kelly 2004: 125-126). As we have seen, a few books have already adopted this approach with interesting results. The aim is not to replace existing approaches and techniques, but to enrich and extend them (cf. Gabrielatos 2005) by finding a place in among them for DDL. This need not be particularly dramatic, as DDL in many respects builds on popular current practices (Boulton 2009c). On the other hand, corpora and DDL need to be given a higher profile within these materials if they are to penetrate public consciousness.

While the emphasis here has been on printed or printable materials as the most familiar and easy-to-use format, another possibility for publishers would be to include corpora of their coursebook documents or of comparable and compatible texts on CD-ROMs or websites which accompany coursebooks, a proposal already mooted by McCarthy (2004: 18). In the case of websites, this does not mean giving something away free to all-comers: firmly anchoring the site to the course represents appreciable publicity, while making the full benefits available only to those who are also in possession of the course itself. Such a measure need not be expensive or difficult to

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

create, as most publishers already have their own software and corpora, although some adaptation might be necessary; copyright should not be a problem for course documents, and searches returning only short extracts may not contravene copyright in any case; and if the CD-ROM or website is planned anyway, the medium itself is not an extra cost. The procedure is not likely to be more difficult or expensive than any other CALL package. Many individuals or small groups have found it possible to provide free access to DDL-style activities and resources on line, such as the multi-media *ELISA*[14] (Braun 2006); on CD-ROM, one might mention *VideoCorpus* (2006), which includes videos along with software to access the transcript corpus, although this is a stand-alone resource and is not accompanied by any activities or suggestions for use (see Nelson 2007, for a brief introduction). While these still encounter some of the objections relating to any ICT activity discussed earlier, anchoring them firmly onto a related coursebook would increase their immediacy and relevance; explicit activities (as opposed to simply making them available for exploration) with appropriate instructions, answers and feedback would make them easier to use at the outset. Facilities like these would represent a substantial pedagogical extra for any course.

## 6. Conclusion

Most of the current interest in DDL is within the research community, with learners in higher education working directly on corpora via a complex interface. Such practices are a worthy goal for those who will need to use a foreign or second language after leaving university; but this is not the case for everyone (cf. Chambers 2005: 114), and hands-on DDL represents a daunting leap for many learners and their teachers, especially in schools. Current research understandably focuses on new things which technology allows learners to do, but in

14 <http://www.uni-tuebingen.de/elisa/html/elisa_info.html>.

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

the process inevitably neglects to consolidate existing gains by comparing results against the reality of ordinary teaching environments. Full hands-on DDL may be possible in research environments, but as Mukherjee (2006: 14) remarks, it is "doubtful […] whether this extremely autonomous corpus-based activity can be fruitfully put into practice in the reality of ELT classrooms." As he goes on to say, this can only happen if teachers are involved in the development of activities and materials in action-research projects, an extremely rare occurrence to date. In other words, researchers should not be surprised that teachers do not listen to them, if *they* do not listen to the teachers (a point forcefully made by McCarthy 2008).

A number of hugely innovative websites have been developed which incorporate DDL in a more guided environment; this certainly makes the approach more accessible, but perhaps does not go far enough. In particular, while ICT can be tremendously motivating for some, it represents an affective barrier for many others, as well as representing a major logistical problem in many environments. Integrating DDL activities into published materials is a natural progression in trying to make it more accessible. It of course represents something of a compromise in an attempt to reconcile the extraordinary (DDL) with the ordinary (published materials). And as with any compromise, it might be necessary to abandon some hard-line principles in order to get others across, to encourage teachers and learners to take the first steps.

DDL itself is not an all-or-nothing process: Gabrielatos (2005) compares the "soft" approach, where teachers lead learners through prepared materials, to the "hard" version of hands-on concordancing; similarly, Mukherjee (2006: 12) refers to a "cline of learner autonomy, ranging from teacher-led and relatively closed concordance-based activities to entirely learner-centred corpus-browsing projects." Even in the case of so-called "deductive DDL" (Cresswell 2007), the learners are still taking an active role in discovering the language, identifying the patterns given and fitting them together. The materials discussed here vary from tenuously DDL to quite staunchly so, suggesting that the approach can be compatible with the printed format at least to some degree. Although this of necessity entails a

certain watering-down of the processes involved, it has been argued here that this is possible without DDL losing its essential characteristics, and that the overall gains outweigh apparent short-term losses.

In particular, the use of published materials can help DDL to reach a wider audience of teachers and learners, forming a key part of the "missionary work" advocated by Römer (2006). It is not the only path available, but one worth pursuing along with other initiatives such as more pedagogically oriented corpora, user-friendly interfaces, and teacher training. Presented in an "ordinary" medium, DDL loses some of its radical image, and thus becomes more amenable to "ordinary teachers and learners in ordinary classrooms" (Mauranen 2004: 208), building as it does on current practices of induction and the use of authentic documents. It can lead to immediate learning, as well as better noticing skills and language awareness which are not necessarily encouraged as part of standard communicative teaching (Carter 1998: 51). Although printed materials do not in themselves promote all the benefits of hands-on DDL, they provide a comparatively accessible lead-in, thus setting the scene for individual exploration later on with the accompanying benefits of greater autonomy, learner-centredness, and life-long learning. Even a small step in this direction is better than no step at all. If O'Keeffe et al. (2007: 247) are right, corpora will become more and more present in coursebooks anyway; better for us as teachers and researchers – and for the learners – if we are involved from the start.



## References

Allan, Rachel 2006. Data-Driven Learning and Vocabulary: Investigating the Use of Concordances with Advanced Learners of English. *Centre for Language and Communication Studies, Occasional Paper,* 66. Dublin: Trinity College Dublin.

Aston, Guy (ed.) 2001. *Learning with Corpora*. Houston: Athelstan.

Alex Boulton. 2010. Data-driven learning: on paper, in practice. In T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, p. 17-52. [pre-publication version]

Barlow, Michael / Burdine, Stephanie 2006. *Phrasal Verbs: American English*. Houston: Athelstan (CorpusLAB).

Bernardini, Sylvia 2001. 'Spoilt for Choice.' A Learner Explores General Language Corpora. In G. Aston (ed.) *Learning with Corpora*. Houston: Athelstan, 220-249.

Boulton, Alex 2008a. But Where's the Proof? The Need for Empirical Evidence for Data-Driven Learning. In M. Edwardes (ed.) *Proceedings of the BAAL Annual Conference 2007*. London: Scitsiugnil Press, 13-16.

Boulton, Alex 2008b. Looking for Empirical Evidence of Data-Driven Learning at Lower Levels. In B. Lewandowska-Tomaszczyk (ed.) *Corpus Linguistics, Computer Tools, and Applications: State of the Art.* Frankfurt: Peter Lang, 581-598.

Boulton, Alex 2009a. Data-Driven Learning: Reasonable Fears and Rational Reassurance. *CALL in Second Language Acquisition: New Approaches for Teaching and Testing. Indian Journal of Applied Linguistics*, 35/1.

Boulton, Alex 2009b. Testing the Limits of Data-Driven Learning: Language Proficiency and Training. *ReCALL*. 21/1, 37-54.

Boulton, Alex 2009c. Data-Driven learning: the perpetual enigma. *Practical Applications in Language and Computers – 7th international conference (PALC 2009)*. Lodz Poland: Uniwersytet Łódski, 6-8 April.

Boulton, Alex Forthcoming. Data-Driven Learning: Taking the Computer out of the Equation. *Language Learning*. 60/3.

Braun, Sabine 2006. ELISA: A Pedagogically Enriched Corpus for Language Learning Purposes. In S. Braun / K. Kohn / J. Mukherjee (eds) *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Frankfurt: Peter Lang, 25-47.

Burdine, Stephanie / Barlow, Michael 2008. *Business Phrasal Verbs and Collocations*. Houston: Athelstan (CorpusLAB).

Capel, Annette 1993. *Concordance Samplers 1: Prepositions.* London: Collins COBUILD.

Carter, Ronald 1998. Orders of Reality: CANCODE, Communication, and Culture. *ELT Journal*. 52/1, 43-56.

Carter, Ronald / McCarthy, Michael (eds) 1988. *Vocabulary and Language Teaching*. London: Longman.

Chambers, Angela 2005. Integrating Corpus Consultation in Language Studies. *Language Learning & Technology*. 9/2, 111-125. <http://llt.msu.edu/vol9num2/pdf/chambers.pdf>.

Chambers, Angela / Kelly, Victoria 2004. Semi-Specialized Corpora of Written French as a Resource in Language Teaching and Learning. *Teanga*. 21, 114-133.

Ciesielska-Ciupek, Maria 2001. Teaching with the Internet and Corpus Materials: Preparation of ELT Materials using the Internet and Corpus Resources. In B. Lewandowska-Tomaszczyk (ed.) *PALC 2001: Practical Applications in Language Corpora.* Frankfurt: Peter Lang, 521-531.

Cobb, Thomas 1997. *From Concord to Lexicon: Development and Test of a Corpus-Based Lexical Tutor.* Montreal: Concordia University, unpublished PhD thesis. <http://www.er.uqam.ca/nobel/r21270/webthesis/Thesis0.html>

Cobb, Tom 2003. Do Corpus-Based Electronic Dictionaries Replace Concordancers? In B. Morrison / G. Green / G. Motteram (eds) *Directions in CALL: Experience, Experiments, Evaluation*. Hong Kong: Polytechnic University, 179-206. <http://www.er.uqam.ca/nobel/r21270/cv/replace_conc.htm>.

Conrad, Susan 2000. Will Corpus Linguistics Revolutionize Grammar Teaching in the 21st Century? *TESOL Quarterly*. 34, 548-560.

Cook, Guy / Seidlhofer, Barbara 1995. An Applied Linguist in Principle and Practice. In G. Cook / B. Seidlhofer (eds) *Principle and Practice in Applied Linguistics: Studies in Honour of H.G. Widdowson*. Oxford: Oxford University Press, 1-25.

Coxhead, Avril 2000. A New Academic Wordlist. *TESOL Quarterly*. 34, 213-238.

Cresswell, Andy 2007. Getting to 'Know' Connectors? Evaluating Data-Driven Learning in a Writing Skills Course. In E. Hidalgo / L. Quereda / J. Santana (eds) *Corpora in the Foreign Language Classroom*. Amsterdam: Rodopi, 267-287.

Estling Vannestål, Maria / Lindquist, Hans 2007. Learning English

Grammar with a Corpus: Experimenting with Concordancing in a University Grammar Course. *ReCALL*. 19/3, 329-350.

Farr, Fiona 2008. Evaluating the Use of Corpus-Based Instruction in a Language Teacher Education Context: Perspectives from the Users. *Language Awareness*. 17/1, 25-43.

Gabrielatos, Costas 2005. Corpora and Language Teaching: Just a Fling or Wedding Bells? *Teaching English as a Second Language – Electronic Journal*. 8/4, 1-35. <http://tesl-ej.org/ej32/a1.html>.

Gaskell, Delian / Cobb, Thomas 2004. Can Learners use Concordance Feedback for Writing Errors? *System*. 32/3, 301-319.

Goodale, Malcolm 1995a. *Concordance Samplers 2: Phrasal Verbs.* London: HarperCollins.

Goodale, Malcolm 1995b. *Concordance Samplers 3: Tenses.* London: HarperCollins.

Granath, Solveig 1998. Using Corpora in Teaching English Syntax to EFL Students at the University Level. In L. Burnard (ed.) *Teaching and Language Corpora 98*. Oxford: Humanities Computing Unit, 87-92.

Hadley, Gregory 2002. Sensing the Winds of Change: An Introduction to Data-Driven Learning. *RELC Journal*. 33/2, 99-124. <http://www.nuis.ac.jp/~hadley/publication/windofchange/windsofchange.htm>.

Hanks, Patrick 2008. The Lexicographical Legacy of John Sinclair. *International Journal of Lexicography*. 21/3, 219-229.

Heyvaert, Liesbet / Laffut, An 2008. Corpora in the Teaching of English in Flemish Secondary Schools: Current Situation and Future Perspectives. In A. Frankenburg-Garcia (ed.) *Proceedings of the 8th Teaching and Language Corpora Conference*. Lisbon, Portugal: Associação de Estudos e de Investigação Científica do ISLA-Lisboa, 400-409.

Higgins, John / Johns, Tim 1984. *Computers in Language Learning*. London: Collins, 88-93.

Jarvis, Huw 2004. Investigating the Classroom Applications of Computers on EFL Courses at Higher Education Institutions in UK. *Journal of English for Academic Purposes*. 3, 111-137.

Johns, Tim 1986. Micro-Concord: A Language Learner's Research Tool. *System*. 14/2, 151-162.

Johns, Tim 1988. Whence and Whither Classroom Concordancing? In P. Bongaerts / P. de Haan / S. Lobbe / H. Wekker (eds) *Computer Applications in Language Learning*. Dordrecht: Foris, 9-27.

Johns, Tim 1991a. Should you be Persuaded: Two Examples of Data-Driven Learning. In T. Johns / P. King (eds) *Classroom Concordancing. English Language Research Journal*. 4, 1-16.

Johns, Tim 1991b. From Printout to Handout: Grammar and Vocabulary Teaching in the Context of Data-Driven Learning. In T. Johns / P. King (eds) *Classroom Concordancing. English Language Research Journal*. 4, 27-45.

Johns, Tim 1997. Kibbitzing One-to-Ones (web notes). *BALEAP: Academic Writing*. University of Reading, 29th November. <http://www.eisu2.bham.ac.uk/johnstf/pimnotes.htm>.

Johns, Tim 2002. Data-Driven Learning: The Perpetual Challenge. In B. Kettemann / G. Marko (eds) *Teaching and Learning by Doing Corpus Analysis*. Amsterdam: Rodopi, 107-117.

Kaltenböck, Gunther / Mehlmauer-Larcher, Barbara 2005. Computer Corpora and the Language Classroom: On the Potential and Limitations of Computer Corpora in Language Teaching. *ReCALL*. 17/1, 65-84.

Koosha, Mansour / Jafarpour, Ali 2006. Data-Driven Learning and Teaching Collocation of Prepositions: The Case of Iranian EFL Adult Learners. *Asian EFL Journal Quarterly*. 8/4, 192-209. <http://www.asian-efl-journal.com/December_2006_EBook.pdf>.

Kosem, Iztok 2008. User-Friendly Corpus Tools for Language Teaching and Learning. In A. Frankenberg-Garcia (ed.) *Proceedings of the 8th Teaching and Language Corpora Conference*. Lisbon, Portugal: Associação de Estudos e de Investigação Cientifíca do ISLA-Lisboa, 183-192.

Lamy, Marie-Noëlle / Klarskov Mortensen, Hans 2007. Using Concordance Programs in the Modern Foreign Languages Classroom. In G. Davies (ed.) *Information and*

*Communications Technology for Language Teachers (ICT4LT)*. Module 2.4. Slough: Thames Valley University. <http://www.ict4lt.org/en/en_mod2-4.htm>.

Leech, Geoffrey 1997. Teaching and Language Corpora: A Convergence. In A. Wichmann / S. Fligelstone / T. McEnery / G. Knowles (eds) *Teaching and Language Corpora*. Harlow: Addison Wesley Longman, 1-23.

Lewis, Michael 1993. *The Lexical Approach: The State of ELT and a Way Forward*. Hove: Language Teaching Publications.

Lewis, Michael 1997. *Implementing the Lexical Approach: Putting Theory into Practice*. Hove: Language Teaching Publications.

Mauranen, Anna 2004. Speech Corpora in the Classroom. In G. Aston / S. Bernardini / D. Stewart (eds) *Corpora and Language Learners*. Amsterdam: John Benjamins, 195-211.

McCarten, Jeanne 2007. *Teaching Vocabulary: Lessons from the Corpus, Lessons for the Classroom*. Cambridge: Cambridge University Press. <http://www.cambridge.org/elt/touchstone/images/pdf/McCarten_booklet.pdf>.

McCarthy, Michael 2004. *Touchstone: From Corpus to Coursebook*. Cambridge: Cambridge University Press. <http://www.cambridge.org/us/esl/Touchstone/teacher/images/pdf/CorpusBookletfinal.pdf>.

McCarthy, Michael 2008. Accessing and Interpreting Corpus Information in the Teacher Education Context. *Language Teaching*. 41/4, 563-574.

McCarthy, Michael / McCarten, Jeanne / Sandiford, Helen 2006. *Touchstone 4*. Teacher's Edition. Cambridge: Cambridge University Press.

Mohamed, Sue / Acklam, Richard 1995. *The Intermediate Choice*. Students' Book and Teacher's Book. Harlow: Longman.

Mukherjee, Joybrato 2006. Corpus Linguistics and Language Pedagogy: The State of the Art – and Beyond. In S. Braun / K. Kohn / J. Mukherjee (eds) *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Frankfurt: Peter Lang, 5-24.

Nation, Paul 1990. *Teaching and Learning Vocabulary*. New York: Newbury House.

Nelson, Mike 2007. A Brief Introduction to Corpora and the Lingonet Video Corpus. <http://www.lingonet.com/web/eng/products/vc1/nelson.htm>.

O'Keeffe, Anne / Farr, Fiona 2003. Using Language Corpora in Initial Teacher Education: Pedagogic, Linguistic and Cultural Insights. *TESOL Quarterly*. 37/3, 389-418.

O'Keeffe, Anne / McCarthy, Michael / Carter, Ronald 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Römer, Ute 2006. Pedagogical Applications of Corpora: Some Reflections on the Current Scope and a Wish List for Future Developments. *Zeitschrift für Anglistik und Amerikanistik*. 54/2, 121-134.

Schmied, J. 2006b. New Ways of Analysing ESL on the WWW with WebCorp and WebPhraseCount. In A. Renouf / A. Kehoe (eds) *The Changing Face of Corpus Linguistics*. Amsterdam: Rodopi, 309-324.

Schmitt, Diane / Schmitt, Norbert 2005. *Focus on Vocabulary: Mastering the Academic Word List*. London: Pearson.

Schmitt, Norbert / McCarthy, Michael (eds) 1997. *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press.

Seidlhofer, Barbara 2000. Operationalizing Intertextuality: Using Learner Corpora for Learning. In L. Burnard / T. McEnery (eds) *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt: Peter Lang, 207-223.

Sinclair, John 2003. *Reading Concordances: An Introduction*. Harlow: Longman.

Sinclair, John / Renouf, Antoinette 1988. A Lexical Syllabus for Language Learning. In R. Carter / M. McCarthy (eds) *Vocabulary and Language Teaching*. Harlow: Longman, 140-158.

Stevens, Vance 1991. Concordance-Based Vocabulary Exercises: A Viable Alternative to Gap-filling. In T. Johns / P. King (eds)

*Classroom Concordancing. English Language Research Journal*. 4, 47-61.

Thompson, Paul 2001. Review of J. Thurstun & C. Candlin: Exploring Academic English. *Language Learning & Technology*. 5/3, 28-31.

Thompson, Paul 2006. Assessing the Contribution of Corpora to EAP Practice. In Z. Kantaridou / I. Papadopoulou / I. Mahili (eds) *Motivation in Learning Language for Specific and Academic Purposes*. Macedonia: University of Macedonia. <http://www.rdg.ac.uk/app_ling/thompson_macedonia.pdf>.

Thornbury, Scott 1995. *Intermediate Choice. Workbook*. Harlow: Longman.

Thornbury, Scott 1999. *How to Teach Grammar*. London: Longman Pearson.

Thornbury, Scott 2004. *Natural Grammar: The Keywords of English and How they Work*. Oxford: Oxford University Press.

Thurstun, Jennifer / Candlin, Christopher 1997. *Exploring Academic English: A Workbook for Student Essay Writing*. Sydney: CELTR.

Thurstun, Jennifer / Candlin, Christopher 1998. Concordancing and the Teaching of the Vocabulary of Academic English. *English for Specific Purposes*. 17/3, 267-280.

Todd, Richard 2001. Induction from Self-Selected Concordances and Self-Correction. *System*. 29/1, 91-102.

Tono, Yukio 2008. TaLC in Action: Recent Innovations in Corpus-Based English Language Teaching in Japan. In A. Frankenberg-Garcia (ed.) *Proceedings of the 8th Teaching and Language Corpora Conference*. Lisbon, Portugal: Associação de Estudos e de Investigação Cientifíca do ISLA-Lisboa.

Tribble, Chris / Jones, Glynn 1997. *Concordances in the Classroom*. 2nd edition. Houston: Athelstan.

Turnbull, Jill / Burston, Jack 1998. Towards Independent Concordance Work for Students: Lessons from a Case Study. *ON-CALL*. 12/2, 10-21. <http://www.cltr.uq.edu.au/oncall/turnbull122.htmll>.

*VideoCorpus* 2006. Turku, Finland: LingoNet.

Warren, Elizabeth 1998. The Use of Concordance Lines in Self-Access Grammar Materials. In L. Burnard (ed.) *Teaching and Language Corpora 98*. Oxford: Humanities Computing Unit, 213-221.

West, Michael 1953. *A General Service List of English Words*. London: Longman.

Whistle, Jeremy 1999. Concordancing with Students using an 'Off-the-Web' Corpus. *ReCALL*. 11/2, 74-80.

Willis, Dave 1990. *The Lexical Syllabus: A New Approach To Language Teaching*. London: Collins.

Willis, Dave 2003. *Rules, Patterns and Words*. Cambridge: Cambridge University Press.

Willis, Jane / Willis, Dave 1988. *Collins COBUILD English Course 2*. London: Collins.