Data-driven learning: the perpetual enigma.

Alex Boulton

Alex Boulton. 2011. Data-driven learning: The perpetual enigma. In S. Goźdź-Roszkowski (ed.), *Explorations across Languages and Corpora*. Frankfurt: Peter Lang, p. 563-580. [pre-publication version]

Data-Driven Learning: The Perpetual Enigma

Alex Boulton
CRAPEL–ATILF/CNRS, Nancy-Université

Abstract

Many uses have been found for corpora in language teaching and learning, most radically perhaps where learners explore the data themselves. Such procedures are especially associated with Tim Johns, who proposed what he called 'data-driven learning' over 20 years ago. While he described his techniques in detail in a number of papers, researchers and practitioners over the years have adapted Johns' procedures and invented new ones of their own, with the result that it can be difficult to pin down exactly what DDL is.

In a tribute to Johns, this paper traces the evolution of DDL through his work from 1984 up until his death in 2009, as well as in DDL studies by other researchers. The enormous variety of activities possible with corpora means it is difficult if not impossible to identify any single element which is either necessary or sufficient for an activity to qualify as DDL. This paper therefore defends a prototype definition of DDL: the further an activity is from the central, prototypical core, the less DDL-like it is, but any cut-off point beyond which we might like to say 'this is no longer DDL' seems likely to be arbitrary rather than empirically grounded or based on a coherent, hermetic definition.

Keywords

authentic, concordancing, corpus, Data-driven learning, DDL, induction, language learning, Tim Johns

Alex Boulton. 2011. Data-driven learning: The perpetual enigma. In S. Goźdź-Roszkowski (ed.), *Explorations across Languages and Corpora*. Frankfurt: Peter Lang, p. 563-580. [pre-publication version]

1. Introduction

In recent decades, electronic corpora have transformed linguistic study almost beyond recognition, in particular enabling vastly improved descriptions of language varieties. In the field of language teaching and learning, a number of researchers (e.g. McCarthy 2008: 564) have hailed this as a truly Kuhnian paradigm shift in many areas, from dictionaries and other resources to syllabus design, course contents and examinations, and even increasing numbers of coursebooks and other materials for both general and specific purposes.

Corpora have multiple potential uses or "affordances", to use the term popularised in applied linguistics by Hafner and Candlin (2007). One of the more controversial of these is as a language resource for learners themselves. This approach is particularly associated with Tim Johns, who started developing it in the 1980s. He first used the term 'data-driven learning' (DDL) in print in the seminal collection of papers co-edited with his colleague, Philip King, in 1991.[1] The introduction has a clear statement:

> All the papers in this issue… describe an application of computers to language-learning that has come to be known as 'classroom concordancing' or 'data-driven learning' (DDL) – the use in the classroom of computer-generated concordances to get students to explore the regularities of patterning in the target language, and the development of activities and exercises based on concordance output. (Johns & King 1991: iii)

Nonetheless, the term remains controversial to this day, with researchers applying the label 'DDL' almost at random to a range of activities. This paper is an attempt to clarify the field and associated concepts, and try to answer the question of what constitutes DDL. We begin by looking at Johns' work to see in greater detail what he intended it to mean, then move on to other studies to see how it has been adopted or rejected in different cases. Finally, we discuss the term itself and its implications for a coherent field of research. The title of this paper is a tribute to Johns' 2002 article, "Data-driven learning: the perpetual challenge", in recognition of the inspiring work of this unceasingly innovative researcher and teacher, who died in March 2009.[2]

2. Background to DDL

The term 'data-driven learning' derives from computer science, describing software which can learn from new data. Because it was Johns who first adopted it to describe the uses of corpora in language learning in 1991, any discussion of DDL will inevitably rely heavily on his work. However, he was not the first to be interested in the concept. McEnery and Wilson (1997: 12) attribute the first applications of corpora to language teaching to Peter Roe at Aston University in 1969, while Johns himself (1986: 159) refers to work by Antoinette Renouf at Birmingham University in the early 1980s using *COBUILD* concordances in the language classroom. The earliest paper Johns (1986) cited along these lines is by Ahmad et al. (1985), who report giving learners access to corpora so they can find multiple examples of language

[1] Actually, one of his papers in the volume (Johns 1991b) which included the term had previously appeared in *CALL Austria* in 1990. Several of his articles were reprinted in various forms, or translated to other languages (Johns 1988b was a French version of Johns 1988a). More interested in teaching than academic research, he was not the most prolific of writers.

[2] Tributes can be found on the homepage of his old centre at the University of Birmingham (http://www.eisu.bham.ac.uk/timjohns.shtml), and in the *International Journal of Corpus Linguistics* (Scott 2009).

phenomena themselves. But a far earlier publication seems to have come from an entirely different source: in 1980, Sandra McKay in San Francisco published an article in *TESOL Quarterly* describing the uses of corpus data for teaching verb usage. The learners in her study are provided with multiple extracts (full sentences) along with instructions to list the types of item that occur before and after the key words, categorising them and observing their behaviour, comparing and contrasting patterns, and so on. DDL thus has a number of roots, coming together in the work of Johns as the main populariser.

Nowhere does Johns provide a single, concrete definition of DDL – or rather, he provides many, defining different aspects and elements of it, along with a tremendous variety of activities. This is inevitable given the multiple affordances of language corpora: Johns and King (1991: iii) describe concordancers as being "as useful a tool in their own way as the Swiss army knife", opening up a "wide range of approaches". Johns himself was certainly not dogmatic about what constituted DDL, and particularly concerned that it should be "integrated with older and more familiar methods" (1991a: 3). Indeed, for him "reasoned eclecticism is a perfectly tenable theoretical position", and any instantiation of DDL (or indeed any other approach) is necessarily "dependent not only on the social, cultural and political setting of a particular society at a particular point in time and the development of education within that setting but also on the technology available in the classroom" (Johns 1988a: 13).

At the University of Birmingham[3] from 1971 until his retirement in 2001, Johns had been in contact with the many researchers working on the *Bank of English* monitor corpus and associated *COBUILD* projects (Sinclair 1987). Working with a concordancer convinced him that it was "one of the most powerful tools that we can offer the language user" (Johns 1988a: 15), inspiring the "classroom use of concordances" which:

> represents rather more than a 'teaching technique' and… can be the basis of a distinctive methodology characterised by the central importance given to the development of the ability of learners to discover things for themselves on the basis of authentic examples of language use. (Johns 1993: 4)

The corpus and concordancer allowed him to "cut out the middleman as far as possible and to give the learner direct access to the data" (Johns 1991b: 30) – one of the most widely quoted phrases he ever wrote. It is therefore largely an inductive approach (Johns 1991b: 29), replacing the traditional 'presentation, practice, production' paradigm (the 'three Ps' so familiar to trainee language teachers in the early communicative era) with a new trilogy which he variously named *identification*, *classification*, *generalisation* (Johns 1991a: 4) or *research*, *practice*, *improvisation* (Johns 1997a: 101). Other researchers have since come up with different labels to characterise the learner's use of language corpora, most famously perhaps the 'three Is' of *illustration*, *interaction*, *induction* (Carter & McCarthy 1995: 155; McEnery et al. 2006: 99).

Clearly aware of the radical shift in the underlying philosophy of language teaching, Johns frequently resorted to metaphors to explain them: "The central metaphors embodying the approach are those of the learner as 'linguistic researcher'[4], testing and revising hypotheses, or as a 'language detective', learning to recognise and interpret clues from context ('Every student a Sherlock Holmes')" (Johns 1997a: 101). This gives rise to another of his widely-quoted passages, that

[3] Specifically the English for Overseas Students Unit (now the English for International Students Unit).
[4] Johns (1988: 13-14) attributes this metaphor to Seliger (1983).

"research is too serious to be left to the researchers"[5] (1991a: 2). Unfortunately, perhaps, the term 'data-driven learning' did little to clarify matters.

3. A hermeneutic approach to data-driven learning

Johns wrote on a number of topics before concentrating on DDL, notably in the fields of reading English for specific or academic purposes, and various CALL applications (Johns 1976, 1980, 1981; Johns & Davies 1983; Johns & Dudley-Evans 1985). Although unconnected to DDL, these papers are helpful in understanding his main concerns in language teaching, with repeated reference to concepts such as induction, problem-solving, puzzling it out, learning by doing, communication and collaboration, as well as the importance of authentic language. A number of later papers are also only tangentially related to DDL, especially describing software (Johns 1996, 1998; Johns & Lixun 1999). In total, Johns' legacy of DDL-related publications resides mainly in nine articles (Johns 1986, 1988a, 1991a, 1991b, 1993, 1997a, 1997b, 2002; Johns et al. 2008), a co-edited collection (Johns & King 1991), a number of on-line collections[6], and a short discussion in a co-edited book (Higgins & Johns 1984). For reasons of space it is not possible to go into details of all his papers here; a summary of each individual paper can be found on the present author's homepage.[7]

Johns briefly referred to the potential uses of corpora in language learning in Higgins and Johns (1984), but it was not until his first major paper on the topic that he gave it any kind of name. This article included the term "classroom concordancing" in the title, but also referred to "data-based learning" in the text (Johns 1988a: 24). "Classroom concordancing" was kept as the title to the 1991 collection of papers. The introduction to this volume also introduced a newer term: "an application of computers to language-learning that has come to be known as 'classroom concordancing' or 'data-driven learning' (DDL)" (Johns & King 1991: iii), with the term 'data-driven learning' also featuring in the titles of both of Johns' own papers in this volume (Johns 1991a, 1991b). If the two terms thus seem to be interchangeable, Johns later (1993: 4) attempted a clarification: "The earlier term Classroom Concordancing described the technique; the new term Data-Driven Learning was coined to emphasise the methodology." However, he was clearly not entirely happy with it: over ten years later (Johns 2002: 107), he referred to it as "an approach… that I have, *for want of a better term*, named data-driven learning" (emphasis added). Finally, in his last paper (Johns et al. 2008: 495) he referred to "corpus-based language learning". While the concepts were presumably clear in his own mind, finding an appropriate label was more difficult: if Johns had his own doubts, it was perhaps not without reason.

Nonetheless, DDL is the term that has stuck in relation to his work, so merits some discussion. The choice of the word **data** is provocative: "what is distinctive about the DDL approach… is the principle that the data is primary" (Johns 1991a: 3).

[5] Based on the French statesman Clémenceau (*Letters and Diaries*, 1935: 6): "La guerre! C'est une chose trop grave pour confier à des militaires." (War is too serious a thing to be left to the military.)

[6] *Virtual DDL Library*. http://www.eisu2.bham.ac.uk/johnstf/ddl_lib.htm;
EAP page. http://www.eisu2.bham.ac.uk/johnstf/timeap3.htm;
Ustí nad Labem DDL Workshop. http://www.eisu2.bham.ac.uk/johnstf/unl_ddl.htm.

[7] Boulton (n.d.). *Tim Johns' Data-Driven Learning: Summaries*, http://arche.univ-nancy2.fr/course/view.php?id=967.

'Data' here crucially refers to corpus data; one might then wonder why he chose 'data' over 'corpus', especially as the former sounds perhaps colder and more scientific. The choice could have been deliberate if he originally wanted to shake others in their complacency, but as corpus linguistics gained ground, he did come back to the more neutral term later (Johns et al. 2008).

The word **driven** is similarly controversial, especially given that he used 'data-*based* learning' in an earlier publications (1988a: 24) and 'corpus-*based* learning' in the last (Johns et al. 2008: 495). The crucial difference from the Birmingham perspective was that a corpus-*driven* approach to language meant trying to put aside all one's previous intuitions and preconceptions in order to look at language afresh only from the evidence of the data, while a corpus-*based* approach meant using corpus data to test existing ideas (cf. Tognini-Bonelli 2001). Whether in language analysis or pedagogy, the term 'driven' is clearly intended to be more radical – the data are primary and everything else has to bend to that; the more nuanced 'based' means taking what one wants from a corpus but not being strait-jacketed by it.[8] As McCarthy and Carter (2003: 338) put it, "the pedagogic process should be informed by the corpus, not driven or controlled by it," although it is difficult to see Johns disagreeing with this position. In the end, one is forced to agree with McEnery et al. (2006: 8) that "the distinction between the corpus-based vs corpus-driven approaches is overstated." Again, we can only presume that Johns' choice was not arbitrary, and in a typically Sinclairean (2004) 'trust the text' frame of mind he wrote of his:

> obstinate belief that even the most refined expert system would here, at best, come a poor second to the teacher's imperfect human system in selecting examples that truthfully reflect the patterning of an item in the corpus, and that students can work with and learn from. (Johns 1997a: 114)

However, it is clear from the many activities he proposed that the teacher's role was crucial in guiding the learner, scaffolding the activities, and mediating the data. He always admitted his print-outs involved "a degree of 'rule-hiding' in the selection of citations, the categories adopted, and the sequencing of citations within each category" (Johns 1991a: 4), and even accepted "editing of concordance data" (Johns 1997a: 114). This is scarcely the radical approach one might expect from the term 'data-driven', and one wonders if the terminological choice was not deliberately provocative: pedagogy would still prevail over ideology.

The concept of **learning** was very much at the centre of Johns' priorities, generating some quite abstract discussion in his 1988a paper. Learning here is thus contrasted with teaching, though this did not mean DDL could not be used "proactively… in a more traditional teacher-centred setting" (Johns 1991b: 31). He was also convinced that DDL would help students to "learn how to learn" (Johns 1991a: 1) and hence to "become better language learners outside the classroom" (Johns 1991b: 31). In many ways, then, the emphasis is less on the language than on the learner. Furthermore, learning was not his only concern, as he also experimented with using corpora as a reference source for translation (2002) or for helping students to correct their own writing (Johns 1986: 61), culminating in his "kibbitzers" (Johns 1997b). Learning may or may not happen, but is incidental and not the main goal when using corpora (or dictionaries, or any other reference resource) in this way for specific inquiries as and when they arise.

[8] Confusingly, McCarthy (e.g. 2008: 566) makes a similar distinction, but with 'corpus-based' as the radical term, and 'corpus-informed' as the more flexible.

4. Empirical research: corpora and language learners

So far we have been looking at what Johns seems to have meant by DDL, on the assumption that, as the inventor of the term in this context, his word is final. That said, the essence of the field of corpus linguistics is to describe common usage and meaning rather than to prescribe it *ex cathedra*. To avoid the trap of etymological fallacy, what it seems to mean to the majority of users has to be taken into consideration.

To identify how the approach has been used by others, this section draws on 61 separate studies looking at some aspect of corpus use for language learners.[9] The definition adopted is necessarily broad, as the aim is to capture the variety of techniques, tasks and procedures used by different researchers, whatever the explicit terminology they use. The survey is limited to empirical evaluations, which tend to provide at least a minimum of background information. The studies were all published in English subsequent to the popularisation of the term DDL in 1991, excluding Johns' own work, and divide into three relatively even-sized groups:

- Type A studies: 19 with a more or less explicit claim to being DDL;
- Type B studies: 23 with at least some reference to Johns' work and/or DDL;
- Type C studies: 19 with no reference at all to Johns or DDL.[10]

The objective then is to compare the three types of study to see what they share and where differences may lie. The first section of Table 1 below provides some background information on the context of the studies. Clearly not all elements are relevant: the fact that the majority of all studies featured English as the target language should not be taken as an indication that English is the 'best' language for DDL (though as a language with comparatively few inflections, it may be more suitable than some). Similarly, the overwhelming majority of studies are conducted among students in higher education, but this is presumably in large part simply because that is where the researchers work rather than due to any inherent limitation of DDL. Other cases are less clear-cut: most studies concern learners at relatively advanced levels of language ability, and it has been argued that DDL is not appropriate at lower levels. Johns (1986: 161) designed *MicroConcord*:

> for a particular type of student (adult: well motivated: a sophisticated learner with experience of research methods in his subject area) with particular needs (fairly closely specifiable in terms of target texts) in a particular learning/teaching situation (in which a great deal of emphasis is placed on developing students' learning strategies and on their responsibility for their own learning).

However, DDL should not be ruled out automatically at lower levels, and Johns (1997a) himself reported using DDL in a 'remedial grammar' course for lower-level international students. A number of other studies do attempt it – especially, as Table

[9] An evolving list of studies can be found on the author's homepage: *Empirical Research in Data-Driven Learning: A Summary*, http://arche.univ-nancy2.fr/course/view.php?id=967.

[10] There is of course no way of knowing whether the lack of reference to Johns' work or DDL means that the authors were unaware of it, or felt it irrelevant to their present paper, or rejected it outright. Similarly, some of the B Studies mention DDL only to dismiss it or suggest an alternative approach. The discussion is intended only to bring out tendencies.

1 shows, among those with a stronger claim to follow Johns' work – and a number of arguments have been put forward (e.g. Boulton 2009a, 2009b).

The most obvious difference between the three types of study lies in the learners' specialisation, with those citing Johns tending to work more with students majoring in areas other than the target language, with an emphasis on English for specific or academic purposes. The importance of this is reflected in the second part of the table, which shows the main focus of the studies: those presenting it as a separate corpus linguistics course; those focusing on one or more specified language areas (e.g. phrasal verbs, connectors, etc.); those using it for written production (including translation and error-correction) or reading; and a small number of others with mixed aims. There are a number of differences between the three types of study, but two stand out: type A studies are more likely to focus on learning ('language'), while type C ones are more frequently interested in corpus use as a reference resource ('writing' or 'reading'). Although these two different uses of corpora rarely receive explicit discussion or differentiation (though see Landure & Boulton [in preparation]), it is possible that some researchers might shy away from claiming to be DDL if their main interest is not learning as such.

		A studies (*n*=19)	B studies (*n*=23)	C studies (*n*=19)	AVERAGE (*n*=61)
background	English	89%	78%	79%	82%
	HEI	89%	87%	89%	88%
	advanced	42%	65%	53%	54%
	intermediate	47%	26%	47%	39%
	low	11%	9%	0%	7%
	specialists	37%	42%	53%	48%
	non-specialists	52%	30%	37%	40%
main focus	corpus linguistics	16%	30%	0%	16%
	language	68%	35%	47%	49%
	writing	0%	22%	42%	21%
	reading	0%	9%	11%	7%
	miscellaneous	16%	4%	0%	8%
support	hands-on	47%	74%	79%	67%
	program	11%	22%	21%	18%
	paper	42%	4%	0%	15%

Table 1. Overview of 61 empirical studies

The final part of the table lists the main type of support or interface used. Type B and C studies are more likely to require learners to interact with corpus data on computer, either directly or via a software package that includes one or more corpora. Type A studies, on the other hand, are more likely to be providing the learners with printed DDL materials. Looking at the table as a whole, this could be because these studies tend to use corpora for learning (with general language items relevant to the entire class) rather than as a resource (with specific language points generated by individual learners), and the learners are probably in greater need of guidance, due both to their lower overall levels of language proficiency and (possibly, given that they are not language specialists) motivation and sophistication.

It is important to bear in mind that these differences reflect general trends only: any of these elements can feature in any of the study types, thus making any hard-

and-fast classification or separation difficult. The same can be said of the corpora and tools involved, and the approaches adopted. Both A and C studies make use of the web as corpus, on-line corpora, locally-built corpora of particular language varieties for specific purposes, large multi-purpose corpora such as the BNC, monolingual and bilingual corpora, and use the same concordancers. Similarly, both types may introduce corpus consultation: as a separate course or integrated to an on-going course; in class, in the computer laboratory or outside; for individual use, pair-work or whole-class activities; proactively in teacher-controlled activities or reactively in response to individual learner queries; using selected samples or an entire corpus; inductively or deductively; in a highly controlled or almost entirely open 'serendipitous' manner; and so on. To summarise: different researchers do different things with corpora, but the differences are at most tendencies, as virtually every type of activity can be found in various studies whether they call themselves DDL or something else.

5. Defining data-driven learning

It is apparent that DDL means different things to different people, sometimes in contradiction with Johns' own work. Virtually the only aspect that is transparently essential in all cases is the use of *corpus* data, and even this is not enough to define DDL. For example, corpora have been used as the basis for course materials such as *Touchstone* (e.g. McCarthy et al. 2006), but the authors' "mediation" makes the corpus origins completely invisible: "Teachers and learners should expect that, in most ways, corpus informed materials will look like traditionally prepared materials" (McCarthy 2004: 15). Few would argue that this is an example of data-driven learning (Boulton forthcoming). DDL may include working with whole texts extracted from a corpus (e.g. Flowerdew 2005), but using whole texts alone is not what most people would understand by DDL – although Johns et al. (2008) based DDL activities on the use of a single novel, *Swallows and Amazons*[11]. DDL is frequently associated with truncated concordance lines or KWICs (key words in context), but may feature whole sentences – as used for example by Schmitt and Schmitt (2005) in their self-proclaimed DDL course *Focus on Vocabulary*. While there are typically several contexts at a time, in some cases there are only two – again, an absolute minimum requirement seems difficult to justify: Johns' examples of the "one item, many contexts" paradigm (e.g. 1997a: 102) frequently featured as few as three lines. In all of these cases, it seems, to qualify as DDL, corpora must not only be present, but must also have a high profile.

The position taken here is that it is simply not possible to identify any element which is both *sufficient* and *necessary* to define an activity as DDL. Certainly it has no monopoly on such apparently key features as induction (Schaffer [1989] presents an intriguing study of induction using full texts, in no way an instantiation of DDL) or authentic materials (recommended extensively in the communicative approach, among others). Watertight definitions are often difficult or impossible to establish in the real world; an alternative is to provide a 'prototype' (Aitchison 2003: ch3), an example which is generally perceived as a good exponent of the class; other instantiations may differ in one or more aspects, but the greater the differences, the less likely they are to be considered members of that class. Gilquin and Gries (2009:

[11] Johns had been working on a biography of the author, Arthur Ransome, during his retirement: http://www.allthingsransome.net/.

6) apply this approach to the concept of a corpus itself, identifying "several criteria that, if met, define a prototypical corpus, but the criteria are neither all necessary nor jointly sufficient."

In the case of data-driven learning, we might propose a prototype such as the following:

- The hands-on use of authentic corpus data (concordances) by advanced, sophisticated foreign or second language learners in higher education for inductive, self-directed language learning of advanced usage.

Yet virtually any part of this description might be absent in a particular case:

- **Hands-on**: Johns (e.g. 1991a, 1991b), like many other researchers, made extensive use of paper-based materials for DDL, and encouraged the use of generic, reusable "ready-made DDL materials" (1991b: 36) to reach a wider audience (as in the examples on his own websites). He even (Johns 1993) mentioned the practice of "blackboard concordancing", getting learners to identify target items from a page of text and write the concordances on the board – no technology required for learners or teachers. Paper-based materials can provide a gentle "way in" (Johns 1993: 7), and "the amount of help on offer can be gradually reduced until students can be presented with concordance output to investigate independently and unaided" (Johns et al. 2008: 495), rather than imposing the new approach (DDL), new materials (corpora), and new technology (software) all at once (Boulton forthcoming). Furthermore, while hands-on DDL may be a long-term goal to promote autonomous corpus investigation for those with strong specific needs (e.g. Johns 1997b), it is not obvious that this is true for all learners.
- **Authentic**: some (notably Widdowson [e.g. 1998]) have claimed that any texts removed from their original context and function should be considered inauthentic in themselves for teaching purposes; even rejecting this extreme position (e.g. Mishan 2004), one might doubt the authenticity of isolated concordance lines chosen by the teacher, especially if they are graded, and possibly even edited, or the use of learner corpora (also envisaged by Johns [1986: 158-159]). Furthermore, despite Johns' repeated insistence on the "overriding importance of authentic text" (Johns & King 1991: iv), corpora of simplified data (e.g. of textbooks, exam papers or simplified novels) might still be used in a DDL approach (e.g. Allan 2009).
- **Corpus data**: some research focuses on whole texts extracted from corpora (e.g. Braun 2005), or a single large text such as a novel (e.g. Johns et al. 2008) or short story (Cobb et al. 2001). As discussed above, the language used also has to be *overtly* extracted from a corpus.
- **Concordances**: in addition to longer texts, or paragraph- or sentence-length extracts, DDL activities may also include study of frequencies, distributions, collocates lists, and so on (e.g. Boulton 2010). The aim is still to generalise from corpus data.
- **Advanced, sophisticated**: the majority of research to date has focused on advanced learners, a common objection being that it "may be all very well for students as intelligent, sophisticated, and well-motivated as ours…, it would not work with students as unintelligent, unsophisticated and poorly-motivated as theirs" (Johns 1991a: 12). But Johns goes on to say, "what I suspect,

however, is that most students given the opportunity to show what they are capable of might be (almost) as remarkable." While noting the prevalence of work with advanced learners, Johns and King (1991: iv) wonder: "How far back towards the beginning could this approach be used?" Yoon and Hirvela (2004) find DDL even more successful with intermediate than with advanced learners, and others report benefits with even lower levels (e.g. Boulton 2008a).

- **Foreign / second language learners**: there seems no obvious reason why DDL should not be applied to learning one's first language, such as in work by Sealey and Thompson (2004). It might also be used by language professionals such as translators who do not consider themselves to be 'learners' as such.
- **Higher education**: occasional studies have been conducted outside formal education (Cobb et al. 2001), independently via the Internet (Smith et al. 2008), in language centres (Allan 2006) and in secondary education, including work by Johns et al. (2008).
- **Inductive**: induction is certainly a central component of DDL (Johns 1991b), but cannot be considered an ultimately defining element. Cresswell (2007) is among those to mention potential deductive uses, comparing corpus data against known or supposed rules to test or refine them, an approach earlier suggested by Johns (1986: 159) himself. However, even in its deductive form it does remain firmly constructivist in outlook, as the learners are not simply given the rules and then told to practise drills.
- **Self-directed**: the freest form of corpus consultation, where learners take on complete responsibility for their learning, has since come to be called "serendipity" learning, most closely associated with Bernardini (2000).[12] But there is "a scale from free to controlled" (Johns & King 1991: iii): for learners at lower levels or with fewer specialist needs, the teacher might usefully retain control over some of the decisions, although in a new guise as "research director and research collaborator rather than transmitter of knowledge" (Johns 1988a: 14). Johns produced CALL software and activities that were "'closed' in the sense that the result is known to the teacher in advance" (Johns 1997a: 101), on the principle that "the more complex the program, the more inaccessible it may become" (Johns 1986: 156), and made extensive use of prepared materials for 'proactive' whole-class use (e.g. 1991a, 1991b).
- **Language learning**: corpus data can also be used as a resource or "informant" (Johns 1991a: 1) as well as explicitly for learning, especially for error-correction or written production in general. Johns mentioned such possibilities early on (1986: 161), developing them into his "kibbitzers" (1993, 1997a) which enable an observer (as in chess [Johns 2002: 111]) to examine the interactions between himself, the concordancer, and learners who came to individual sessions for help with their writing. He developed a parallel concordancer specifically for DDL work for translation as well as for learning (Johns 1993: 7), and made use of translation for "reciprocal learning" (Johns 2002), where learners are paired so that each is a native speaker of the other's target language and collaborate to work on parallel corpora, as well as

[12] Johns had himself used the expression "serendipity learning", but referring to a specific activity where pairs of learners are provided with print-outs of different concordances "with the instruction to find as many things as possible in ten minutes that are worth reporting back to the rest of the class" (Johns 1988: 21).

in his final paper (Johns et al. 2008). It can also be used for testing purposes, particularly in the form of multiple blanked concordances, whether on paper (e.g. Stevens 1991) or in CALL software (e.g. Johns 1997a).

- **Advanced usage**: DDL is often associated with "what is normally untaught and possibly unteachable" (Johns 1991b: 28), on the "'collocational border' between syntax and lexis… [which is where] DDL methods seem to be most effective" (Johns 2002: 109). But DDL has also been used to help with skills work (mainly writing and reading), and the activities proposed in Johns (1991b) include making sentences from fragments for increased coherence, inferring meaning from context, guessing the background situation, constructing new dialogues, and so on. It has frequently been used for vocabulary learning (e.g. Cobb 1999), occasionally for wider grammatical patterns (e.g. Johns 1997a; Boulton 2008b), and even pronunciation (Gut 2006).

To allow for all these exceptions, the only hard-and-fast definition would have to be something like 'language users exploiting language corpora', so vague as to be nearly useless – hence the relevance of a prototype definition. One important corollary of all this is that DDL is not an all-or-nothing affair: its boundaries are fuzzy, and any identifiable cut-off point will necessarily be arbitrary. With first-time users, especially given lower levels of language proficiency, sophistication and motivation, it might well be appropriate to start with a relatively 'weak' form of DDL (near the boundary), gradually moving on to a more 'hard-core' version (at the prototypical centre) as and when necessary.

In many cases, the teacher is likely to remain the centre of attention at the start, deciding the language points to study, the corpora and tools to use, the activities, progression, and so on; as the learners become familiar with all of this, they can gradually take on more responsibility for their own learning by deciding these issues for themselves. Such a gradual changeover from a proactive to a reactive approach is truly 'autonomising' in the classic sense intended by Holec (e.g. 1981). It might be easier to adopt a deductive approach initially, testing rules or hypotheses against the data rather than trying to work them out unaided. Carefully-designed activities with known outcomes can give way to more open-ended exploration, especially as learners move from fixed materials (especially print-outs) to hands-on corpus consultation. This can also be a way to make the data more accessible as an alternative to simplifying the corpus itself, as concordance lines can be selected, graded and organised in advance.

6. Conclusion

Johns frequently referred to DDL as an 'approach', outlining a number of associated techniques; it is not a 'method' in its own right, but can be integrated into various types of courses each with their own methods. In their classic textbook, Richards and Rodgers (2001: 19) make the following distinctions:

> *Approach* is the level at which assumptions and beliefs about language and language learning are specified; *method* is the level at which theory is put into practice and at which choices are made about the particular skills to be taught, the content to be taught, and the order in which the content will be presented; *technique* is the level at which classroom procedures are described. [emphasis added]

In other words, an approach does not give rise to "a specific set of prescriptions and techniques" but "a variety of interpretations as to how the principles can be applied", meaning it can be "revised and updated over time" (Richards & Rodgers 2001: 245). The spirit inherent in the DDL approach is clear: empowering learners to explore language corpora and come to their own conclusions. An enormous variety of techniques can be used within this approach: the absence of any can not be taken to mean that an activity is not DDL – precisely because it is not a method and avoids the dogmatism associated with the level of method. Such criticism is therefore not only unfounded, but also does a disservice to the field as a whole.

One might wonder whether the name one applies is of any real importance, but it has been argued here that the term 'data-driven learning' may in itself be off-putting for some: 'data' sounds cold and clinical (whereas DDL is highly learner-centred), 'driven' sounds uncompromising (whereas DDL is anything but dogmatic), and 'learning' is only one aspect (DDL can also refer to corpus use as a resource for writing or translation, for example). If no label is generally accepted, then the approach as a whole cannot assume a coherent identity, essential if the "'trickle down' from research to teaching" is to turn into the torrent confidently predicted by Leech over a decade ago (1997: 2). A readily-identifiable brand name helps to increase marketability in any sphere, and language pedagogy is no exception: without it, research is perceived as fragmented, reducing its overall visibility. 'Data-driven learning' might or might not be the phrase to take the field further, only time will tell. Other proposals have included '(classroom) concordancing', 'corpus consultation', and 'corpus-based language learning' among others, each also with its relative merits and disadvantages.

This is not to denigrate the potential of the approach itself. Johns (1993: 8) claimed it "mirror[ed] the Zeitgeist of the 1990s", but it fits just as well with the post-communicative movement of the present day which, even in caricature, "turns the school into a house of personal learning and discovery, task-based, collaborative, with process-input, the teacher as guide, and the like" (Decoo 2001: n.p.). The eclectic and non-dogmatic nature of DDL (by whatever name) is fully compatible with communicative language teaching; discovery learning and learning by doing; autonomisation and learning to learn; learner-centredness and individualization; collaborative learning and creativity; task-based and process as well as product orientations; form and meaning in constructivism; with an emphasis on the authentic language of discourse by register / genre. It can be particularly useful in language for specific purposes (including content and language integrated learning), making full use of information and communication technology and thus promoting computer skills.

It seems reasonable then to hope for a bright future where DDL can integrate ordinary classroom practice.

Bibliography

Ahmad, K., G., Corbett & M. Rogers. (1985), 'Using computers with advanced language learners: an example.' *The Language Teacher* (Tokyo), 9 (3): 4-7.

Aitchison, J. (2003), *Words in the Mind: An Introduction to the Mental Lexicon*, Oxford: Blackwell.

Allan, R. (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.

Allan, R. (2009), 'Can a graded reader corpus provide "authentic" input?' *ELT Journal*, 63: 23-32.

Bernardini, S. (2000), 'Systematising serendipity: proposals for concordancing large corpora with language learners.' In: L. Burnard & T. McEnery (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt: Peter Lang. 225-234.

Boulton, A. (2008a), 'Looking for empirical evidence for DDL at lower levels.' In: B. Lewandowska-Tomaszczyk (Ed.), *Corpus Linguistics, Computer Tools, and Applications – State of the Art*, Frankfurt: Peter Lang/ 581-598.

Boulton, A. (2008b), 'DDL is in the details… and in the big themes.' In: M. Davies, P. Rayson, S. Hunston & P. Danielsson (Eds.), *Proceedings of 4th Corpus Linguistics Conference*. http://ucrel.lancs.ac.uk/publications/CL2007/

Boulton, A. (2009a), 'Testing the limits of data-driven learning: language proficiency and training.' *ReCALL*, 21 (1): 37-51.

Boulton, A. (2009b), '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): 81-106.

Boulton, A. (2010), 'Data-driven learning: taking the computer out of the equation.' *Language Learning*, 60 (3).

Boulton, A. (forthcoming), 'Data-driven learning: on paper, in practice.' In: T. Harris & M. Moreno Jaén (Eds.), *Corpus Linguistics in Language Teaching*, Bern: Peter Lang.

Braun, S. (2005), 'From pedagogically relevant corpora to authentic language learning contents.' *ReCALL*, 17 (1): 47-64.

Carter, R. & M. McCarthy. (1995),' Grammar and the spoken language.' *Applied Linguistics*, 16: 141-158.

Cobb, T. (1999), 'Breadth and depth of lexical acquisition with hands-on concordancing.' *Computer Assisted Language Learning*, 12 (4): 345-360.

Cobb, T., C. Greaves & M. Horst. (2001), 'Can the rate of lexical acquisition from reading be increased? An experiment in reading French with a suite of on-line resources.' In: P. Raymond & C. Cornaire (Eds.) *Regards sur la Didactique des Langues Secondes*. Montréal: Editions Logique. 133-135. http://www.er.uqam.ca/nobel/r21270/cv/BouleE.htm

Cresswell, A. (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.

Decoo, W. (2001), 'On the mortality of language learning methods.' L. Barker lecture. Brigham Young University, 8 November. http://www.didascalia.be/mortality.htm

Flowerdew, L. (2005), 'An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: countering claims against corpus-based methodologies.' *English for Specific Purposes*, 24 (3): 321-332.

Gilquin, G. & S. Gries. (2009), 'Corpora and experimental methods: a state-of-the-art review.' *Corpus Linguistics and Linguistic Theory*, 5(1): 1-26.

Gut, U. (2006), 'Learner speech corpora in language teaching.' In: S. Braun, K. Kohn & J. Mukherjee (Eds.), *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, Frankfurt: Peter Lang. 69-86.

Hafner, C. & C. Candlin. (2007), 'Corpus tools as an affordance to learning in professional legal education.' *Journal of English for Academic Purposes*, 6 (4): 303-318.

Higgins, J. & T. Johns. (1984), *Computers in Language Learning*, London: Collins.

Holec, H. (1981), *Autonomy in Foreign Language Learning*, Oxford: Oxford University Press.

Johns, T. (1976), 'The communicative approach to language teaching in the framework of a programme of English for academic purposes.' In: E. Roulet & H. Holec (Eds.), *L'Enseignement de la Compétence de Communication en Langues Secondes*, Neuchâtel: Institut de Linguistique de l'Université de Neuchâtel. 94-112.

Johns, T. (1980), 'The text and its message: an approach to the teaching of reading comprehension for students of development in administration.' In: H. Faber & A. Maley (Eds.), *Leseverstehen im Fremdensprachenunterricht*, Munich: Goethe Institut.

Johns, T. (1981), 'The uses of an analytic generator: the computer as teacher of English for specific purposes.' *ESP Teacher: Role, Development and Prospects. ELT Documents*, 112. London: British Council.

Johns, T. (1986), 'Micro-Concord: a language learner's research tool.' *System,* 14 (2): 151-162.

Johns, T. (1988a), '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, T. (1988b), 'Implications et applications des logiciels de concordance dans la salle de classe.' *Les Langues Modernes*, 82 (5): 29-45.

Johns, T. (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, T. (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, T. (1993), 'Data-driven learning: an update.' *TELL&CALL,* 2: 4-10.

Johns, T. (1996), 'If our descriptions of language are to be accurate… A footnote to Kettemann.' *TELL&CALL,* 4: 44-46. http://www.eisu2.bham.ac.uk/johnstf/ifbeto.htm

Johns, T. (1997a), 'Contexts: the background, development and trialling of a concordance-based CALL program.' In: A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and Language Corpora.* Harlow: Addison Wesley Longman. 100-115.

Johns, T. (1997b), 'Kibbitzing one-to-ones.' Paper prepared for the *BALEAP meeting on Academic Writing.* University of Reading, 29 November. http://www.eisu2.bham.ac.uk/johnstf/pimnotes.htm

Johns, T. (1998), *Multiconcord: The Lingua Multilingual Parallel Concordancer for Windows.* http://www.eisu.bham.ac.uk/johnstf/lingua.htm

Johns, T. (2002), 'Data-driven learning: the perpetual challenge.' In: B. Kettemann & G. Marko (Eds.), *Teaching and Learning by Doing Corpus Analysis.* Amsterdam: Rodopi. 107-117.

Johns, T. & F. Davies. (1983), 'Text as a vehicle for information: the classroom teaching of reading in ESP.' *Reading in a Foreign Language*, 1 (1): 1-19. http://nflrc.hawaii.edu/rfl/PastIssues/rfl11johns.pdf

Johns, T. & A. Dudley-Evans. (1985), 'An experiment in team-teaching of overseas postgraduate students of transportation and plant biology.' In: J. Swales (Ed.), *Episodes in ESP*. Oxford: Pergamon.

Johns, T., L. Hsingchin & W. Lixun. (2008), 'Integrating corpus-based CALL programs in teaching English through children's literature.' *Computer Assisted Language Learning*, 21 (5): 483 – 506.

Johns, T. & P. King (Eds.), (1991), *Classroom Concordancing. English Language Research Journal,* 4.

Johns, T. & W. Lixun. (1999), 'Four versions of a sentence-shuffling program.' *System,* 27 (3): 329-338.

Landure, C. & A. Boulton. (in preparation), 'Corpus et autocorrection pour l'apprentissage des langues.' *ASp*.

Leech, G. (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.

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

McCarthy, M. (2008), 'Accessing and interpreting corpus information in the teacher education context.' *Language Teaching*, 41 (4): 563-574.

McCarthy, M. & R. Carter. (2003), 'What constitutes a basic spoken vocabulary?' *Cambridge University ESOL Examinations Research Notes*, 13: 5-7.

McCarthy, M., J. McCarten & H. Sandiford. (2006), *Touchstone 4* (teacher's edition), Cambridge: Cambridge University Press.

McEnery, T. & A. Wilson. (1997), 'Teaching and language corpora.' *ReCALL*, 9 (1): 5-14.

McEnery, T., R. Xiao & Y. Tono. (2006), *Corpus-Based Language Studies: An Advanced Resource Book*, London: Routledge.

McKay, S. (1980), 'Teaching the syntactic, semantic and pragmatic dimensions of verbs.' *TESOL Quarterly*, 14 (1): 17-26.

Mishan, F. (2004), 'Authenticating corpora for language learning: a problem and its resolution.' *ELT Journal*, 58 (3): 219-227.

Richards, J. & T. Rodgers. (2001), *Approaches and Methods in Language Teaching* (2nd edition), Cambridge: Cambridge University Press.

Schaffer, C. (1989), 'A comparison of inductive and deductive approaches to teaching foreign languages.' *Modern Language Journal*, 73: 395-403.

Schmitt, D. & N. Schmitt. (2005), *Focus on Vocabulary: Mastering the Academic Word List*, London: Pearson.

Scott, M. (2009), 'In memory of Tim Johns.' *International Journal of Corpus Linguistics*, 14 (3): 271-274.

Sealey, A. & P. Thompson. (2004), 'What do you call the dull words? Primary school children using corpus-based approaches to learn about language.' *English in Education*, 38 (1): 80-91.

Seliger, H. (1983), 'The language learner as linguist: of metaphors and realities.' *Applied Linguistics*, 4 (3): 179-191.

Sinclair, J. (1987), *Looking Up: An Account of the COBUILD Project in Lexical Computing*, London: Collins.

Sinclair, J. (2004), *Trust the Text: Language, Corpus and Discourse*, London: Routledge.

Smith, S., A. Chen & A. Kilgarriff. (2008), 'A corpus query tool for SLA: learning Mandarin with the help of SketchEngine.' In: B. Lewandowska-Tomaszczyk (Ed.) *Corpus Linguistics, Computer Tools, and Applications – State of the Art*, Frankfurt: Peter Lang. 673-686.

Stevens, V. (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.

Tognini-Bonelli, E. (2001), *Corpus Linguistics at Work*, Amsterdam: Benjamins.

Widdowson, H. (1998), 'Context, community, and authentic language.' *TESOL Quarterly*, 32 (4): 705-716.

Yoon, H. & A. Hirvela. (2004), 'ESL student attitudes toward corpus use in L2.' *Journal of Second Language Writing*, 13 (4): 257-283.