Corpora in language teaching and learning

Alex Boulton

Research Timeline

Corpora in language teaching and learning

Alex Boulton

Atilf – University of Lorraine & CNRS, France

alex.boulton@univ-lorraine.fr

Crapel/Atilf, University of Lorraine, BP 13397, 54015 Nancy cedex, France

ALEX BOULTON is Professor of English and Applied Linguistics at the University of Lorraine. His research interests have long been in lexis, language learning and new technologies, and where these come together in corpus linguistics and its uses for 'ordinary' teachers and learners – DATA-DRIVEN LEARNING. He has published and edited books and papers in these fields over the years, and is on various boards and committees, including for TaLC (Teaching and Language Corpora) and *ReCALL* (co-editor). He is currently head of the Crapel (Didactique des Langues et Sociolinguistique), a research group in applied linguistics at the Atilf (University of Lorraine & CNRS).

This timeline looks at explicit uses of corpora in foreign or second language (L2) teaching and learning, i.e. what happens when end-users explore corpus data, whether directly via concordancers or integrated into CALL programs, or indirectly with prepared printed materials. The underlying rationale is that such contact provides the massive contextualised exposure needed for language learning, but in a more controlled way than purely haphazard exposure via regular reading or listening, thus promoting or enhancing noticing, language awareness, autonomy, and ultimately producing 'better learners'. It interweaves with many other notions in language teaching, from authenticity and autonomy to induction and constructivism, taking in much of what we know about language from corpus linguistics itself – its fuzzy, probabilistic nature and the importance of lexical patterns, collocations, chunks, frequencies and distributions. Commonly known as DATA-DRIVEN LEARNING or DDL (Johns 1990), the approach goes back several decades; yet despite its impressive pedigree, a frequent lament is that it remains a marginal practice. Teachers may not be aware of DDL if it is absent from their initial training, or may see it as a research activity confined to higher education. While corpus consultation may appear too demanding for school-age learners, there are connections with the web searches that are already a frequent practice for many. Though lack of uptake is sometimes attributed at least in part to a dearth of empirical research, this claim is increasingly difficult to defend, as witnessed by the number of entries in this timeline and the hundreds of others not included. For a balanced discussion of the limits of DDL as well as what it can reasonably achieve, see Kaltenböck & Mehlmauer-Larcher (2005)[1].

The goal here has been to provide an overview of the evolution of the field from the earliest publications up until the present day. The first objective was to focus on empirical evaluations of DDL based on an extensive trawl of empirical research which identified over 200 publications. In some cases, the choice is quite straightforward, especially where a name is associated with a single widely-cited and influential paper; in others, it is difficult to choose among an author's cumulative output of a dozen or more publications which could almost merit an individual timeline of their own. Few authors have more than one entry here as the aim is to provide as broad a sweep as possible. On occasion, an intermediate solution was adopted, with the same authors contributing a base study along with a follow-up or replication study in the same entry. As a failsafe, the reference sections of the various papers were analysed to identify authors of more than one empirical DDL study; the 25 most frequently cited are all included here.

This procedure on its own would not do justice to the full range of activity in the field, and some key players would be excluded since their most widely cited work is non-empirical. Since research is a fundamentally human activity, a second consideration was to focus on the individuals and their most inspirational work. Empirical research is thus complemented with a number of descriptive, argumentative or position papers, especially from the early days, along with a handful of more theoretical works that provide some foundation to DDL that would otherwise be missing, as well as the main surveys reported in the area. Only manuals and textbooks were ultimately excluded from the timeline itself, though some of these provide additional theoretical support for the approach; selected references for further reading are provided as supplementary materials on the *Language Teaching* website. Some difficult decisions still needed to be made, and though the final list inevitably has an element of subjective choice, some of my personal favourites have had to be omitted for reasons of space and coherence according to the various inclusion criteria.

DDL draws on many different strands of research, and some of the most frequently cited names from neighbouring fields could only be included in a heterogeneous timeline of several hundred entries. These range from corpus linguistics itself and the development of corpora, tools and methodologies that are widely used in DDL, including work on language for specific purposes, frequency lists and learner corpora, error-correction and contrastive analysis, to corpus uses in syllabus design, materials and resources, testing and translation, as well as fields from lexis to acquisition studies and more general language learning theory.

While much work has been reported in a few major journals (notably *Computer Assisted Language Learning*, *ReCALL*, *Language Learning & Technology*, *System* and *CALICO Journal*), the bulk appears as chapters in various collections, often arising from conferences or in less well-known journals or conference proceedings, often available on line. In favouring the most accessible and prestigious sources here, the result is inevitably biased towards publications in and on English and from certain regions of the world, and does not address reports in other languages, or using corpora or software specific to other languages.

Ultimately, the timeline papers fall into four main categories as outlined below. Since there is no overarching programme for DDL research, the balance is not perfect between the various categories, which often overlap.

A **Theoretical underpinnings**. As mentioned above, this category does not attempt to outline all the general language teaching and learning theory that the approach draws on (e.g. cognitive skills or autonomisation), and is limited to those landmark papers most directly connected to DDL itself. Much of this derives from research in corpus linguistics, which has influenced our understanding of language (e.g. in the nature of lexicogrammar, patterning or chunks).

B **Descriptive papers**. A potentially enormous category, so largely limited here to a few influential papers from the relatively early days.

C **Empirical evaluations** are divided into sub-sections.

First, mainly emic papers, often featuring questionnaires, interviews, logs etc., which focus on:

C1 learners' attitudes and representations about corpus use;

C2 learners' behaviour and processes in corpus use.

Second, largely etic papers, often with a pre/post-test or control/experimental design from tests or analyses of learner productions, which look at:

C3 effects of corpus use as a learning aid;

C4 effects of corpus use as a reference resource.

D **Surveys and syntheses**. This category includes papers that attempt to make sense of the increasing body of research in the field.

Several features emerge. In chronological terms, it is in the nature of the developmental curve in applied linguistics research that a new field is developed by a small number of visionaries first in descriptive and speculative papers (beginning around 1990), followed by increasing numbers of ever more focused and sophisticated empirical evaluations (especially from 2000 on, with a dozen or more such papers every year since 2005) until such time as enough are available for various types of research syntheses, with theoretical underpinnings appearing at various points along the way. Curiously, perhaps, while many empirical studies refer to theoretical and pedagogical foundations, few seek directly to test them, and theory has not been a major driving force leading to new practices. The main initial impetus came largely from different spheres in the UK (Birmingham, Lancaster and Nottingham, among others), but it has since widened into much of the rest of the world: other English-speaking countries, western Europe, the Middle East and Asia in particular. The relative paucity of DDL work in the United States of America, however, would bear further examination.

The methodologies have become increasingly rigorous and sophisticated in design – not to mention complex in the case of the many PhDs not covered here – with a range of data collection instruments. Interestingly, there remains a balance of emic and etic studies with questionnaires ranking alongside pre/post-tests or control/experimental designs as the instruments of choice. This allows considerable insight not just into the outcomes but also into the participants' representations and the processes involved. There has been a move over time from purely lexicogrammatical to more discourse-oriented studies, and from use of corpora as a learning aid towards their use as a reference resource, especially in language for specific or academic purposes, with a predictable dominance of English as the target language. Some issues prove controversial, such as the need for training, or the potential with lower proficiency levels or for oral skills. More longitudinal work is needed outside university classrooms: in autonomous contexts, in secondary education, in language schools, or in the workplace for 'genuine' life-long needs; as well as the possibilities of exploiting more everyday tools for DDL-compatible web searches.

YEAR	REFERENCE	ANNOTATION	THEME
1980	McKay, S. (1980). Teaching the syntactic, semantic and pragmatic dimensions of verbs. *TESOL Quarterly* 14.1, 17–26.	Corpora had been used informally in language teaching for some time, a use associated with Johns at Birmingham, but the first published report is by **McKay** in San Francisco. The hypothesis sets the stage for much of DDL – that exposure to repeated patterns of authentic use can lead to learning, inductively or deductively. The paper describes various activities for paper-based materials to encourage noticing from corpus-derived examples (a crucial concept not exclusive to DDL; see Schmidt 1990)[2], followed by corpus-based exercises and productive activities.	**B**
1986	Johns, T. (1986). Micro-Concord: A language learner's research tool. *System* 14.2, 151–162.	Though **Johns** had made earlier brief mention of possible uses of corpora in language learning, this is his first full-length paper on the topic. Here he presents his MicroConcord, which – with Mike Scott. – later evolved into the ever-popular WordSmith Tools. Originally conceived specifically as a 'language learner's research tool', Johns was keen for it to remain simple to avoid overwhelming the learner with functions and data. Various features are described and potential pedagogical activities suggested for learning and error-correction.	**B**
1989	Baten, L., A.-M. Cornu & L. K. Engels (1989). The use of concordances in vocabulary acquisition. In C. Laurent & M. Nordman (eds.), *Special language: From humans thinking to thinking machines*. Clevedon: Multilingual Matters, 452–467.	Surprisingly undercited, the first empirical evaluation of DDL exhibits considerable ecology and complexity, allowing 400 Dutch economics students to select and explore their own texts in conjunction with the Brown corpus for English according to their own interests over four months. Their written work and a prepared oral exam showed remarkable sophistication in corpus use. Feedback is overwhelmingly positive regarding motivation and the possibilities for individualisation and autonomy in contextualised vocabulary work; but **Baten et al.** also note negative points relating mainly to the limitations of the corpus and availability of the concordancer.	**C1, C2**
1990	Johns, T. (1990). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *CALL Austria* 10, 14–34.	One of **Johns**' most frequently cited papers, especially in the 1991 version in Johns & King (eds.)[3] along with its sister paper there ('Should you be persuaded: Two samples of data-driven learning materials'), this contains the first occurrence of the term 'data-driven learning', variously referred to elsewhere as 'classroom concordancing' or 'corpus-based learning'. Given the technological limitations of the time, much of his early work concentrated on use of generic or 'proactive' printed materials derived from corpora. Johns describes the approach as 'revolutionary', promoting the authenticity of corpus data and the role of learner as researcher with direct access to that data rather than being dependent on the teacher. Various activities are discussed, with examples in the appendix.	**A, B**

1991	Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.	A tremendously influential figure in moving linguistics back towards empirical investigation, **Sinclair** is often considered the father of modern corpus linguistics. Founder of the Cobuild project in Birmingham (cf. JOHNS 1990), his research was mainly descriptive/lexicographical but had a practical interest in learning and teaching. This widely-cited book presents the case for combining lexis and grammar, suggesting that language use is based on the 'idiom principle' as opposed to the traditional teaching emphasis on grammar, such that learning happens by deconstructing chunks of input which are later used to build up utterances. This is supported by recent psycholinguistic evidence beyond the scope of this timeline (e.g. Millar 2011)[4].	**A**
1991	Stevens, V. (1991). Concordance-based vocabulary exercises: A viable alternative to gap-fillers. In T. Johns & P. King (eds.), *Classroom concordancing. English Language Research Journal* 4, 47–61.	The only paper that goes beyond description or suggestion in this seminal volume edited by Johns & King, **Stevens** tests students' lexicogrammar in a traditional gap-fill text or multiple concordance lines. Each participant served as control and experimental participant, completing half the questions using traditional gap-fills, half using concordances. The concordance questions derived from a corpus of the students' textbooks were answered significantly better in the second session, suggesting the importance of training.	**C3**
1993	Fligelstone, S. (1993). Some reflections on the question of teaching, from a corpus linguistics perspective. *ICAME Journal* 17, 97–109.	This paper is widely cited (see Leech 1997 for a follow-up)[5] for its three-way characterisation of corpora in teaching. **Fligelstone** first noted 'teaching about' corpus linguistics as an academic subject in its own right, and 'teaching to exploit' corpora in giving students training in the mechanics of hands-on corpus use for various questions. Most importantly here, 'exploiting to teach' included both indirect applications (i.e. selection and sequencing of items to teach in accordance with corpus findings) and hands-on concordancing for language learning purposes – i.e. DDL.	**A**
1994	Ma, B. K. C. (1994). Learning strategies in ESP classroom concordancing: An initial investigation into data-driven learning. In L. Flowerdew & A. Tong (eds.), *Entering text*. Hong Kong: Hong Kong University of Science and Technology Language Centre, 197–214.	This study mainly relies on a specially compiled corpus of chapters from computer manuals for a 10-hour computing course in Hong Kong to help correct written work. Students' search queries are analysed and compared against questionnaires and interviews with randomly selected participants. Despite some success, **Ma**'s main conclusion is that, left to their own devices, students do not always use their time efficiently, formulating inappropriate queries and underusing some functions, and concentrate on corpus use for error-correction only.	**C1, C2**

1997	Aston, G. (1997). Involving learners in developing learning methods: Exploiting text corpora in self-access. In P. Benson & P. Voller (eds.), *Autonomy and independence in language learning*. London: Longman, 204–214.	**Aston**'s initial attempt at DDL with his advanced Italian learners of English was only partly successful as the tasks were too open (cf. MA 1994), and learners tended to focus on isolated examples and miss patterns and tendencies. A second class provided greater focus with a small corpus of medical research articles where learners were encouraged to look first for non-linguistic information (i.e. content). Working in small groups and still with a substantial degree of autonomy, they successfully pursued their own questions in browsing mode, reporting their findings clearly to others.	**C2**
1997, 1999	Cobb, T. (1997). Is there any measurable learning from hands-on concordancing? *System* 25.3, 301–315. Cobb, T. (1999). Applying constructivism: A test for the learner-as-scientist. *Educational Technology Research & Development* 47.3, 15–33.	These two studies derive from **Cobb**'s Ph.D. using his own software, since developed into the highly popular Compleat Lexical Tutor (http://www.lextutor.ca/). The software included corpus data, the objective being to learn hundreds of high-frequency words in a single year. Though control groups using word lists and dictionaries were successful at acquiring definitional knowledge in the short term, the experimental group did significantly better in retaining it long-term and in transferring it to new texts, and rated the materials highly.	**A, C1, C3**
1997	McEnery, T. & A. Wilson (1997). Teaching and language corpora. *ReCALL* 9.1, 5–14.	Best known for their corpus linguistics work, **McEnery & Wilson** had considerable interest in early DDL, and their influence continues to this day. This paper outlines key concepts from autonomy, discovery learning and individualisation to self-assessment, correction and translation, with the teacher as guide (cf. JOHNS 1990). It is often cited for the 'percolation' of corpus data and methodologies into language learning and its call for more focused empirical research, as well as a reprise of the trichotomy: learning about corpora, teaching to exploit corpora, and exploiting to teach with corpora. It also argues for pedagogical corpora, and the need for DDL to break out of higher education.	**A, B**
1998	Thurstun, J. & C. N. Candlin (1998). Concordancing and the teaching of the vocabulary of academic English. *English for Specific Purposes* 17.3, 267–280.	Frequently cited for their innovative workbook published the previous year, **Thurstun & Candlin** here provide the rationale underpinning it. Designed for university students requiring academic English, the book presents a small number of rhetorically related items using overtly corpus-derived materials following a four-stage process akin to 'the three Ps' but initially inductive: look, familiarize, practise, create (see also Carter & McCarthy 1995[6] for 'the three Is': illustration, interaction, induction). The obvious objection is the amount of work involved for so few target items, but the authors defend this as exposing learners to far more language through concordance contexts, and promoting deep processing and cognitive skills which should have longer-term benefits.	**A, B**

1998	Turnbull, J. & J. Burston (1998). Towards independent concordance work for students: Lessons from a case study. *ON-CALL* 12.2, 10–21.	Case studies allow more in-depth analysis of learners' concordancing activities, as reported by **Turnbull & Burston** for two students enrolled for a master's degree in Australia. Radically different motivations and learning styles meant that one (field independent) student used the concordancer frequently and effectively, and felt she learned a great deal, while the other (field dependent) showed less interest and was less successful, and found it a waste of time. It is perhaps surprising that more work has not been conducted relating DDL and learning styles, strategies and motivations.	**C1, C2**
2000	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.	**Bernardini**'s take on autonomous corpus browsing has become an accepted definition, popular especially for corpus-driven post-graduate academic writing needs. Here she attempts to balance it with greater guidance in a 'pedagogy of discovery' (e.g. ASTON 1997). Third-year Italian translation students were required to analyse agreed questions on their own out of class; reporting back, their conclusions were generally found to be correct, and their enthusiasm is clear as they indicated using the BNC for other courses too. Problems are noted, such as concentration on the unusual, frustration in formulating search strings, and underuse of some software functions.	**C1, C2**
2000	Seidlhofer, B. (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.	Following a disappointing first attempt with relatively 'technophobic' students, **Seidlhofer** aimed to increase motivation and involvement by asking her English students in Austria to examine corpora of their own writing using WordSmith Tools. Despite the small corpus of very short texts, learners were able to discover many of the elements of corpus linguistics and compare their texts against larger corpora. Allowing learners greater personal involvement in the corpus and in the questions to explore led to far greater enthusiasm and motivation (cf. ASTON 1997).	**B**
2001, 2010	Kennedy, C. & T. Miceli (2001). An evaluation of intermediate students' approaches to corpus investigation. *Language Learning & Technology*, 5.3, 77–90. Kennedy, C. & T. Miceli (2010). Corpus-assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference resource. *Language Learning & Technology* 14.1, 28–44.	These two studies have language students in Australia consult the authors' own corpus of Italian, with detailed analysis of individual student behaviours. The 2001 students proceeded from controlled activities, to correcting their own work, to exploring their own questions; most found it useful, though difficulties are described in the main stages of formulating the question, devising a search strategy, observing and selecting examples, and drawing conclusions. The 2010 learners used the corpus and bilingual dictionaries to revise their own creative writing. **Kennedy & Miceli**'s characterisation of 'pattern hunting' (looking for content or language ideas) vs. 'pattern defining' (checking usage) has achieved widespread recognition. The results, along with interviews and questionnaires, show diverse levels of uptake, and successful corpus use is linked to trial and error and ownership with further training as the main way forward.	**C1, C2**

2001	Todd, R. W. (2001). Induction from self-selected concordances and self-correction. *System* 29.1, 91–102.	Some difficulties in DDL relate to the unfamiliar nature of the technology and how to use it; **Todd** attempts to circumvent these by using regular search engines to provide 'concordances' from the web. Two lexical errors were highlighted in the work of each of 23 postgraduate students in Thailand; most successfully derived 'rules' from their web searches and produced valid corrections. The most difficult items were those with multiple meanings or patterns of use. For an update on DDL-like use of the web as concordancer, see Boulton (2015).[7]	**C4**
2003	Lee, C.-Y. & H.-C. Liou (2003). A study of using web concordancing for English vocabulary learning in a Taiwanese high school context. *English Teaching and Learning* 27.3, 35–56.	**Lee & Liou**'s 17-year-old high-school students spent 10 weekly sessions using DDL for vocabulary, with a variety of corpus activities based around their textbook contents. All three levels improved in the post-test, with the differences between them decreasing; this may suggest that DDL can be accessible at lower levels. Learners with an inductive preference tended to perform better with DDL, and even those with a deductive profile were quite enthusiastic despite alleged cultural orientation towards knowledge-transmission in Taiwan, where considerable DDL work has been successfully carried out.	**C1, C3**
2003	Sun, Y.-C. & L.-Y. Wang (2003). Concordancers in the EFL classroom: Cognitive approaches and collocation difficulty. *Computer Assisted Language Learning* 16.1, 83–94.	Again in Taiwan, **Sun & Wang** had their high-school students study collocations inductively from concordances, or deductively using traditional methods. While the control group was provided with rules, the experimental group searched for the target patterns to arrive at their own rules. The learners corrected errors in sentences before and after the experiment, with the experimental group showing greater improvement, especially for the apparently easier items.	**C3**
2004, 2006	Chambers, A. & Í. O'Sullivan (2004). Corpus consultation and advanced learners' writing skills in French. *ReCALL* 16.1, 158–172. O'Sullivan, Í. & A. Chambers (2006). Learners' writing skills in French: Corpus consultation and learner evaluation. *Journal of Second Language Writing* 15.1, 49–68.	**Chambers & O'Sullivan** looked at Irish learners of French in postgraduate (2004) and undergraduate (2006) programmes. In both studies, the learners first wrote an essay in the L2 with access to traditional resources. Following training in corpus use, they were allowed two hours to use concordances to correct their original texts, changes being marked as positive, negative or no effect. Improvement was noted in all areas, even for items which could have been more easily checked in dictionaries; reactions were positive overall, with most claiming they would use corpora in the future.	**C1, C4**

2004	Curado Fuentes, A. (2004). The use of corpora and IT in evaluating oral task competence for tourism English. *CALICO Journal* 22.1, 5–22.	Most corpus work features written texts; **Curado Fuentes** is notable for his use of corpora of transcribed speech, here for academic and professional texts with upper-intermediate students in tourism. Over two weeks, the course moved from identifying recurrent forms in short texts to online concordancing and related activities, and was evaluated via prepared and spontaneous spoken tasks. The experimental group showed considerably more effective use of items that occurred in the corpus in the prepared task; they also exhibited more effective use of cohesive markers in the spontaneous discussion, and claimed greater confidence. Though the experimental group made as many errors, this is attributed to their overall faster pace of delivery and longer presentations.	**C3, C4**
2004	Mukherjee, J. (2004). Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany. In U. Connor & T. Upton (eds.), *Applied corpus linguistics: A multidimensional perspective*. Amsterdam: Rodopi, 239–250.	Arguing that the teacher's perspective is crucial in promoting DDL, **Mukherjee** reports on a series of four-hour workshops in Germany. In all, 248 experienced English teachers enrolled; questionnaires show that they overwhelmingly intend to use corpora in the future, but mainly see the relevance for themselves as learners and, as teachers, for preparing paper-based DDL materials, with only 12% expecting to have their learners consult corpora directly. This is perhaps not surprising in a secondary school context, and age was found to be a factor with younger teachers more willing to experiment. However, it does suggest some limitations to the fully autonomous, serendipitous type of corpus consultation often equated with DDL (cf. BERNARDINI 2000).	**C1**
2004	Yoon, H. & A. Hirvela (2004). ESL student attitudes toward corpus use in L2. *Journal of Second Language Writing* 13.4, 257–283.	Much work in DDL has focused not on outcomes but on learner beliefs and attitudes to corpus use, a trend crystallised in this paper by **Yoon & Hirvela**. They introduced corpora to their learners in the US over several weeks, providing feedback on writing assignments conducted out of class. Detailed questionnaires show considerable enthusiasm, with the overwhelming majority claiming they would use corpora in the future, though the authors follow up negative reactions too. The paper is also notable for its practical approach and bringing DDL to lower proficiency levels (cf. LEE & LIOU 2003), who seemed more favourable than the advanced group, perhaps as they had received more guidance.	**C1, C2**
2005	Braun, S. (2005). From pedagogically relevant corpora to authentic language learning contents. *ReCALL* 17.1, 47–64.	Written corpora are comparatively easy to collect, while many learners may have greater need of oral skills (cf. CURADO FUENTES 2004). A series of papers by **Braun** is important for its focus on multimedia corpora, beginning with ELISA – a small collection of interviews which can be watched individually or searched through a concordancer aligning text with sound and video, and are enriched with word lists etc. The corpus is specifically designed with pedagogical rather than corpus-linguistic aims in mind (cf. MCENERY & WILSON 1997). Small-scale testing with secondary-school learners features in later papers.	**B**

2005	Chan, T.-P. & H.-C. Liou (2005). Effects of web-based concordancing instruction on EFL students' learning of verb-noun collocations. *Computer Assisted Language Learning* 18.3, 231–251.	Much work with Asian learners has recourse to parallel corpora, often in conjunction with the authors' own software and related CALL materials as in this paper by **Chan & Liou**. Their first-year EFL course covered verb-noun collocates either inductively using a corpus, or deductively without, via various online activities. An immediate post-test showed significantly greater improvement for collocates via DDL than in the control group, an advantage maintained in a delayed post-test 10 weeks later; it is also argued that an inductive approach takes time to produce its maximum effect. There was no significant correlation between proficiency level and post-test scores (cf. LEE & LIOU 2003); a questionnaire showed generally favourable attitudes to DDL, which did correlate with performance.	**C1, C3**
2005	Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.	**Hoey**'s lexical priming draws on usage-based theories in claiming that grammar is essentially the product of innumerable encounters with words in context, thus following SINCLAIR's (1991) focus on lexicogrammar. Mastery of a language (L1 or L2) is mastery of primings (i.e. expectations of meaning and use in context), and learning crucially consists of repeated exposure to items in context; the role of teaching is to fast-track the process so the learner can notice different uses.	**A**
2005	Kaur, J. & V. Hegelheimer (2005). ESL students' use of concordance in the transfer of academic word knowledge: An exploratory study. *Computer Assisted Language Learning* 18.4, 287–310.	While vocabulary lists have their uses, they also have their limitations (cf. COBB 1997/1999), as highlighted in this study on academic vocabulary by **Kaur & Hegelheimer**. Their ESL students in the US were introduced to concordancing over two weeks, and used the BNC via LexTutor for an out-of-class writing assignment in addition to dictionaries. While no significant difference was found in terms of overall performance (attributed to the time limitations in this experimental study), the DDL students did use the target items significantly more frequently and more accurately in the written assignment.	**C1, C2, C3**
2006	Lee, D. & J. Swales (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes* 25.1, 56–75.	**Lee & Swales** join forces to bring corpus-based discourse/genre work to the classroom; this paper is particularly cited for its detailed notes on a 15-week course designed to help research students write academic papers. Work covered a number of language points decided by the teachers, though learners compiled corpora of their own and had freedom of choice in the final project, presented as a conference paper. The students were largely successful, appreciating their increased autonomy with language specific to their needs; this is reflected in the fact that most of them bought WordSmith Tools for their own future use.	**B, C1, C2**
2007	Chambers, A. (2007). Popularising corpus consultation by language learners and teachers. In E. Hidalgo, L. Quereda & J. Santana (eds.), *Corpora in the foreign language classroom*. Amsterdam: Rodopi, 3–16.	**Chambers** provides one of the earliest surveys of empirical research in DDL. Though she covers only 12 studies, she manages to paint a broad picture of research up to that time, noting that it is mostly conducted by practising teachers/researchers in an individual in-class experimental approach; research outside the classroom is almost non-existent. The studies are broadly action-research in design, with greater reporting of qualitative results (learner attitudes and representations, processes and behaviours) than quantified learning outcomes.	**D**

2007	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.	The question addressed by **Hafner & Candlin** is what students would make of a corpus on their own initiative. The first cohort had just a brief demonstration supplemented by online tutorials, but with disappointing uptake. The introduction was expanded to a full hour in the second year, doubling the number of students who used the corpus to 40%. Interviews with selected students showed that this may still not be enough: some 'adopters' made little use of the corpus for its intended linguistic function, but browsed it for its legal content – an alternative 'affordance'.	**C1, C2**
2008	Johns, T. F., H. Lee & L. Wang (2008). Integrating corpus-based CALL programs and teaching English through children's literature. *Computer Assisted Language Learning* 21.5, 483–506.	In **Johns**' final paper, with colleagues in Taiwan, the 'corpus' consists of a single novel to be read traditionally by a control group and accompanied by various CALL tools and DDL activities in the experimental group. Questionnaires showed very positive attitudes and greater motivation in the experimental group; they also performed significantly better than the control in the post-test for reading comprehension, and read twice as fast. End-of-term exam results suggest improvement extending beyond the specific tasks covered – a tantalising discovery and one sorely in need of following up, since a major claim of DDL is that it can help students become better learners.	**C1, C3**
2009	Gilmore, A. (2009). Using online corpora to develop students' writing skills. *ELT Journal* 63.4, 363–372.	Corpora are often used to help with writing, as in **Gilmore**'s study with lower-level Japanese learners of English for academic purposes. Following a brief introduction to large on-line corpora, they then worked out of class to improve items marked as wrong on an earlier assignment. The outcomes were rated by four judges, who found the revisions produced more 'natural' text in nearly two thirds of cases; feedback from the learners was also overwhelmingly positive.	**C1, C4**
2010	Boulton, A. (2010a). Data-driven learning: Taking the computer out of the equation. *Language Learning* 60.3, 534–572.	While many studies have suggested extensive training is crucial (e.g. KENNEDY & MICELI 2001/2010), **Boulton** shows here that even lower-level learners are capable of detecting patterns in carefully prepared corpus-derived data after only a five-minute introduction. As with LEE & LIOU (2003), DDL was particularly useful for students at lower levels of proficiency , while the more advanced learners maintained their advantage using the traditional approach. A questionnaire completed after the experimental condition showed very positive reactions.	**C1, C3**
2010	Boulton, A. (2010b). Learning outcomes from corpus consultation. In M. Moreno Jaén, F. Serrano Valverde & M. Calzada Pérez (eds.), *Exploring new paths in language pedagogy: Lexis and corpus-based language teaching*. London: Equinox, 129–144.	Despite frequent claims that DDL lacks empirical support, **Boulton** here reports on 27 evaluations of learning outcomes from DDL. The weight of evidence is positive overall, even if the conclusion at this stage is that the variety of research designs and questions would make a formal meta-analysis virtually impossible (but cf. COBB & BOULTON 2015). The 'real' benefits of DDL may be only partially assessable by short-term experimental studies; more longitudinal studies of cognitive development in particular are called for.	**D**

2010	Conroy, M. A. (2010). Internet tools for language learning: University students taking control of their writing. *Australasian Journal of Educational Technology* 26.6, 861–882.	This study introduced concordancing alongside advanced Google searches, on the assumption that the two involve similar techniques. The students at this Australian university found the training useful, especially concordancers. Nearly half already used the internet for language learning; while the initial figures were far lower for corpus work, they increased dramatically by the end of the course. A case study of three learners found them subsequently making more and better use of Google in error-correction (cf. TODD 2001), which **Conroy** takes to mean that accessibility, familiarity and ease of use are crucial, with a promising future for 'Google-driven language learning'.	**C1, C2, C4**
2010	Park, K. & C. Kinginger (2010). Writing/thinking in real time: Digital video and corpus query analysis. *Language Learning & Technology* 14.3, 31–50.	This case study (cf. TURNBULL & BURSTON 1998) had a single advanced learner in the US using a corpus for writing. **Park & Kinginger** track all corpus queries and record the computer screen; the participant later reviews these and comments. The results suggest that the writing process is divided into 'transactions' on specific questions, a fifth of which are complex in requiring several related queries. This study is important in providing an insight to the cognitive processes involved in corpus consultation.	**C1, C2**
2011, 2012	Pérez-Paredes, P., M. Sánchez-Tornel, J. M. Alcaraz Calero & P. Aguada Jiménez (2011). Tracking learners' actual uses of corpora: Guided vs non-guided corpus consultation. *Computer Assisted Language Learning* 24.3, 233–253. Pérez-Paredes, P., M. Sánchez-Tornel & J. M. Alcaraz Calero (2012). Learners' search patterns during corpus-based focus-on-form activities: A study on hands-on concordancing. *International Journal of Corpus Linguistics* 17.4, 483–515.	**Pérez-Paredes** and colleagues in Spain have been notable for rigorous tracking of how learners use corpora instead of relying solely on emic feedback. The 2011 study provided corpus guidance for the experimental group, who were found to perform better but also to make greater use of other internet resources and complete more activities, despite having less time available. Though the 2012 study found many queries were relatively unsophisticated, it also reports a combination of corpus and other online searches producing the most successful outcomes. These papers further highlight the comparability of corpus and internet searches (cf. CONROY 2010), and suggest that concordance designers have much to learn from search engines.	**C2**

2012, 2014	Charles, M. (2012). Proper vocabulary and juicy collocations: EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes* 31.2, 93–102. Charles, M. (2014). Getting the corpus habit: EAP students' long-term use of personal corpora. *English for Specific Purposes* 35.1, 30–40.	Graduate students of mixed language and disciplinary backgrounds who need English for academic writing feature in several studies by **Charles** in the UK. The 2012 paper outlines how students create their own disciplinary corpora of research articles for analysis using AntConc; the 50 questionnaires reveal considerable enthusiasm and use outside class, and highlight potential problems. The 2014 paper provides a rare look at continued usage one year on; of the 40 respondents, 86% used their corpora in some way (38% regularly) for writing or revising their academic writing. Though the initial course had focused on discourse, most continued use was for lexicogrammar.	**C1, C2**
2012	Chujo, K. & K. Oghigian (2012). DDL for EFL beginners: A report on student gains and views on paper-based concordancing and the role of L1. In J. Thomas & A. Boulton (eds.), *Input, process and product: Developments in teaching and language corpora*. Brno: Masaryk University Press, 170–183.	Work on English in Japan is well represented by a series of papers by **Chujo** & **Oghigian** and their colleagues, who are particularly interested in the use of DDL at lower levels of proficiency (cf. BOULTON 2010a). This paper has their engineering students use a parallel news corpus to study vocabulary, noun phrases and verb phrases over two semesters, mainly hands-on. Pre- and post-tests show the experimental group making significant gains in almost all areas, while the control group made very limited improvement. The following year, a comparable group did most activities on paper; a mix of paper-based and hands-on DDL is therefore recommended.	**C3**
2012, 2014	Frankenberg-Garcia, A. (2012). Learners' use of corpus examples. *International Journal of Lexicography* 25.3, 273–296. Frankenberg-Garcia, A. (2014). The use of corpus examples for language comprehension and production. *ReCALL* 26.2, 128–146.	These two experimental studies by **Frankenberg-Garcia** compare the ability of Portuguese learners of English to derive information from single or multiple (3-line) corpus examples vs dictionary definitions (cf. BOULTON 2010a) at university level (2012) and in secondary education (2014). Neither study found a significant difference between multiple examples and dictionary definitions in comprehension, but multiple examples were significantly more helpful in production. The 2014 paper selected different examples for receptive and productive purposes, giving still greater advantage to multiple examples on traditional problem items, with implications for lexicography as well as teaching.	**C4**

2012	Taylor, J. R. (2012). *The mental corpus: How language is represented in the mind.* Oxford: Oxford University Press. Boulton, A. & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning* 67.2.	Taking HOEY (2005) a step further, **Taylor** claims that language in the mind is itself analogous to a mental corpus in terms of input, processing and output. Knowledge of a language is the cumulative experience of multiple encounters with different items, with repeated patterns reinforcing each other in normal usage. This accounts for the probabilistic rather than rule-generated nature of language, with its inherent fuzziness and messiness. Learning can be promoted by focus on exposure in context with noticing of exemplars, in line with usage-based models.	**A**
2013	Geluso, J. (2013). Phraseology and frequency of occurrence on the web: Native speakers' perceptions of Google-informed second language writing. *Computer Assisted Language Learning* 26.2, 144–157.	Learners' frequent use of Google as a language aid (cf. PÉREZ-PAREDES et al. 2011, 2012) is exploited here as a route into DDL with only a brief introduction and three short sessions. **Geluso** had 25 Japanese students search for formulaic sequences using double quotation marks to improve self-selected 'questionable phrases' in their own writing, revising according to frequency. Using a similar procedure to GILMORE (2009), four raters compared the 167 phrases before and after, finding a significant increase in 'naturalness'.	**C4**
2014	Chang, J.-Y. (2014). The use of general and specialized corpora as reference sources for academic English writing: A case study. *ReCALL* 26.2, 243–259.	**Chang**'s study has two main objectives: to compare a general and a specialised corpus, and to see how learners use them in their own time for genuine writing needs. Ten master's and doctoral students with varying degrees of publication experience used the corpora over 22 weeks; one-to-one advising sessions were recorded and transcribed, and interview data collected. Both types of corpora were found to be useful, sometimes in different ways: the general corpus (COCA) was more 'credible' due to its size and inclusion mainly of native-speaker data, while the specialised corpus of research articles was considered more relevant.	**C1, C2**
2014	Yoon, H. & J. W. Jo (2014). Direct and indirect access to corpora: An exploratory case study comparing students' error correction and learning strategy use in L2 writing. *Language Learning & Technology* 18.1, 96–117.	In a rare study comparing use of printed corpus data and hands-on concordancing (cf. CHUJO & OGHIGIAN 2012), **Yoon & Jo** report a case study with Korean learners exposed to five weeks of each treatment. The same writing task performed before and after the course showed substantial improvement in accuracy rates. The correction rate was higher in the indirect use, but the participants appreciated the direct use more and appreciated formulating their own searches. They also used more (and more different) strategies working hands-on; these are discussed in detail based on existing taxonomies.	**C1, C2, C3**

2015	Cobb, T. & A. Boulton (2015). Classroom applications of corpus analysis. In D. Biber & R. Reppen (eds.), *The Cambridge handbook of English corpus linguistics*. Cambridge: Cambridge University Press, 478–497.	These papers take up BOULTON's (2010) challenge and include the first attempts at meta-analysing of the outcomes of L2 corpus use, both reporting substantial effects. In the second, more extensive paper, **Boulton & Cobb** were able to meta-analyse 88 unique samples from 64 studies out of the initial pool of 205 DDL publications. The results show a mean gain (pre-/post-test) effect size of 1.50 and a mean difference (control/experimental post-test) effect size of .95; both figures are in the top quartile of meta-analyses in second language acquisition as a whole (Plonsky & Oswald 2014),[8] showing that DDL can have substantial impact. An analysis of potential moderator variables challenges some of the traditional arguments about where DDL does and does not work.	**D**
2015	Tribble, C. (2015). Teaching and language corpora: Perspectives from a personal journey. In A. Leńko-Szymańska & A. Boulton (eds.), *Multiple affordances of language corpora for data-driven learning*. Amsterdam: John Benjamins, 37–62.	**Tribble** is well placed to offer an overview of how the field has developed over the decades. In addition to a personal retrospective, of interest here is the latest in his series of surveys into corpus use. While the respondents are partially self-selecting, this nevertheless provides a useful profile of the types of people who use corpora in their teaching, the corpora and tools they use, and the purposes they put them to for themselves and for their students. Analysis of open-ended responses provides deeper insights into the respondents' perceptions of corpus use, its advantages and disadvantages.	**D**

[1] Kaltenböck, G. & B. Mehlmauer-Larcher (2005). Computer corpora and the language classroom: On the potential and limitations of computer corpora in language teaching. *ReCALL* 17.1, 65–84.
[2] Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics* 11.2, 129–158.
[3] Johns, T. & P. King (eds.) (1991). *Classroom concordancing. English Language Research Journal* 4.
[4] Millar, N. (2011). The processing of malformed formulaic language. *Applied Linguistics* 32.2, 129–148.
[5] 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.
[6] Carter, R. & M. McCarthy (1995). Grammar and the spoken language. *Applied Linguistics* 16.2, 141–158.
[7] Boulton, A. (2015). Applying data-driven learning to the web. In A. Leńko-Szymańska & A. Boulton (eds.), *Multiple affordances of language corpora for data-driven learning*. Amsterdam: John Benjamins, 267–295.
[8] Plonsky, L. & F. L. Oswald. (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning* 64.4, 878–912.