



HAL
open science

Applying data-driven learning to the web

Alex Boulton

► **To cite this version:**

Alex Boulton. Applying data-driven learning to the web. Agnieszka Lenko-Szymanska & Alex Boulton. Multiple Affordances of Language Corpora for Data-driven Learning, John Benjamins, pp.267-295, 2015, 10.1075/scl.69.13bou . hal-00937993

HAL Id: hal-00937993

<https://hal.science/hal-00937993>

Submitted on 7 Aug 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Alex Boulton. (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, p. 267-295. DOI: 10.1075/scl.69.13bou

Abstract

Data-driven learning typically involves the use of dedicated concordancers to explore linguistic corpora, which may require significant training if the technology is not to be an obstacle for teacher and learner alike. One possibility is to begin not with corpus or concordancer, but to find parallels with what ‘ordinary’ users already do. This paper compares the web to a corpus, regular search engines to concordancers, and the techniques used in web searches to data-driven learning. It also examines previous studies which exploit web searches in ways not incompatible with a DDL approach.

Keywords

data-driven learning; DDL; web as corpus; Internet; search engines; Google

1. Introduction

Data-driven learning (DDL) involves the use of dedicated concordancers to explore large language corpora. Or does it? The basic concept is commonly associated with Tim Johns who coined the term in 1990. Though he certainly was mainly concerned with corpora and concordancers available at Birmingham University, some of which he created himself, his various writings leave the concept open to much wider interpretation. As the technology is itself a frequently cited difficulty, one possibility is to begin not with a corpus or a concordancer, but with the learners and what they already do. In what ordinary, everyday activities outside the language classroom are learners involved in using computers to search for information? Most obviously, in browsing the web.¹ The temptation then is to wonder whether the web could serve as a substitute corpus, and Google or another search engine as a substitute concordancer.² This might seem inappropriately iconoclastic to some, but DDL and corpus linguistics have broken more than one ideological barrier in their time. If these resources had been available in the 1980s, corpus linguistics and DDL would likely be very different today.

While recognising that we have here neither a corpus nor a concordancer in their traditional senses, the essential point of contact is that Google + web provides a means to explore huge collections of language data, and the failings noted by some linguists may not be as relevant in language teaching/learning as they are for research purposes – they may even be considered advantages in some cases. The main point is that the criteria are simply different for the two communities (Stewart *et al.* 2004). In using corpora for pedagogical purposes, it is important not simply to apply corpus linguistics in the language classroom (cf. Widdowson 2000 on “linguistics applied”). For this, pedagogical uses need to “emerge from the shadow

¹ Though there is a technical difference between the *Internet* and the *world wide web*, the two are commonly used interchangeably, as here.

² Google is the main example given in this paper as it is the most widely used search engine in the world today, easily beating Bing and Yahoo! as well as other engines less well known in western countries, such as Baidu and Yandex (see <http://comscore.com> for recent figures). No value judgement is intended on the relative advantages or appropriateness of any individual search engine.

of Corpus Linguistics and demonstrate that the uses of corpora do not have to be restricted to the ways in which (corpus) linguists like to use them” (Braun 2010: 92).

There are two basic issues we need to consider. Firstly, if DDL does not absolutely require a corpus or a concordancer, then there would seem to be no essential reason not to consider the web and Google for possible use in DDL; this is the topic of Section 2. Secondly, if DDL does absolutely require them, then the question becomes whether the web is sufficiently similar to a corpus, and Google sufficiently similar to a concordancer; these are the topics of Sections 3 and 4 respectively. The paper then explores some of the functions of Google as they relate to learning needs in Section 5, and rounds off in Section 6 with a survey of recent research which involves learners exploring the web with search engines in DDL-like ways.

2. Data-driven learning in theory and practice

The basic methodology of DDL involves L2 users exploring the target language for themselves rather than ‘being taught’. This is generally equated with access to a corpus via a concordancer, though the benefits attributed to DDL may apply more generally to learner engagement with any type of language data. Johns did choose the term *data-driven* (as opposed to corpus-driven), and discussed many other types of data, from a single novel (*Swallows and Amazons*: Johns et al. 2008) to *ad hoc* collections of highly diverse text types and genres, e.g. serious and popular science articles alongside lecture transcripts and newspaper articles, and even some fiction “to give an occasional glimpse of sex-and-violence” (Johns 1997: 104). In Johns’ vision, therefore, DDL does not seem to rely essentially on a corpus as it is understood in corpus linguistics – i.e. a large collection of authentic texts in electronic format, designed to be representative of a language variety (e.g. Biber et al. 1998; Sinclair 2005; McEnery et al. 2006; Cheng 2011). What is important about the language data is that it should be pedagogically relevant (cf. Braun 2005), and that the learner should engage directly with that data rather than relying on the teacher as intermediary (Boulton & Tyne 2014). Any advantages that accrue to DDL in its canonical form apply to all learner exploration of language.

Despite his insistence that “the data is primary” (1991: 3), in most cases Johns seems in fact more concerned with the *processes* involved. The first of his assumptions about corpus use is that “the learner’s engagement with text should play a central role in the learning process” (Johns 1988: 10), giving rise to a methodology where “central importance [is] 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). In this way:

The central metaphors embodying the approach are those of the learner as ‘linguistic researcher’, testing and revising hypotheses, or as a ‘language detective’, learning to recognise and interpret clues from context (‘Every student a Sherlock Holmes’). (Johns 1997: 101)

DDL can thus be defined as “using the tools and techniques of corpus linguistics for pedagogical purposes” (Gilquin & Granger 2010: 359) – a broad but elegant definition of DDL which notably does not mention ‘corpora’ *per se*. The obvious tool is the concordancer, though the label is potentially misleading as it can do far more than mere concordancing: Johns and King (1991: iii) likened it to a Swiss army knife, and the range of functions has increased considerably in the intervening years. In addition to KWIC concordances, full

sentences and longer contexts (potentially useful in discourse analysis), many concordancers also provide information about the frequency and distribution of particular items, charts, tables and plots of statistical information, frequency lists of words or lemmas (if the corpus is lemmatised), n-grams, collocates, keywords, and so on (cf. Charles, this volume). Whether such functions are performed by a concordancer or more specialised software, they may be included under the umbrella term of DDL as long as they allow the learner to engage with the language data. The term ‘concordancer’ is thus more a convenient shorthand than an accurate description of most corpus software.

In the end, we are left with the conclusion that there are fuzzy boundaries to DDL; while some may insist that it only applies to hands-on concordance work with a traditional concordancer, others may interpret it more widely (Boulton 2011a). In either case, it would seem desirable to encourage learners’ capacity for critical thinking about language through hypothesis formation and testing, from imagining how to turn real questions into exploitable queries, to sorting, analysing and interpreting the data. However, the expected DDL “‘trickle down’ from research to teaching” (Leech 1997: 2) has by and large not materialised, and corpus consultation remains a marginal activity in most language learning contexts. A major reason for this may be real or perceived technological obstacles relating to the uses of corpora or associated software (e.g. Rodgers et al. 2011; Philip 2011). It may simply be asking too much of the learner to deal with these technological demands since they are already having to get to grips with new processes of sophisticated linguistic reasoning. Putting corpus and software together may demand quite sophisticated linguistic reasoning if the learner is to act as a “researcher” (Johns 1988), identifying relevant language points, formulating hypotheses about them, transforming these into appropriate queries, interpreting the results, and refining the process until a usable outcome is obtained. Interpreting corpus data can be quite different from ‘normal’ reading of text (Sinclair 2003 devotes an entire book to Reading Concordances), and substantial training may be required to master the technical aspects (e.g. Leńko-Szymańska 2014). And “for a number of teachers (and learners), the technology behind DDL may... be too difficult, even if they are given a basic training” (Römer 2010: 30). In sum, “it is unsurprising that learners find it difficult to get to grips with new material (the corpora), new technology (the software) and a new approach (DDL) all at once” (Boulton 2010a: 539).

Of course, this all depends on how DDL is implemented, and several ways of reducing the difficulties can be envisaged. First, in a more teacher-fronted version, corpus data can be used in printed materials designed to reduce the difficulties both of the presentation of the language and of the tasks to be carried out. Johns (1990: 19) made considerable use of such “proactive” materials, and clearly counted them as DDL since they still involve the learner exploring the language data. He also referred positively to Willis’s (1998) blackboard concordancing as an instantiation of DDL (Johns 1993), even though it involved no more than providing each small group of learners with a separate page of printed text, and asking them to identify a given feature (e.g. prepositions) and to copy it out on the board in the usual KWIC format. These types of ‘hands-off’ procedure (i.e. not interacting with the software; see Boulton 2012) may lead to immediate learning benefits and help to increase language awareness in the long term. But however useful they are, they depend on materials and input designed by others and which the learner has little or no control over; and when these

are unavailable after class, the learner will no longer be able to continue. Further, many of the decisions will be taken out of the hands of the individual learner, relying on generic materials which may not be equally relevant to all. Removing the technology entirely may be appropriate in some contexts for some learners, as I have argued elsewhere (Boulton 2010b, 2012), but will necessarily involve a loss of some of the advantages of DDL.

A second possibility is to keep the tools but to bring them closer to the learner, rather than expecting the learner to adjust entirely to the tools of conventional corpus linguistics. In this respect, a number of attempts have indeed been made to introduce more pedagogically friendly corpora – small corpora (Aston 1997), genre-specific corpora (Ghadessy et al. 2001), multimedia corpora (Braun 2005), learner corpora (Cotos 2014), self-compiled corpora (Charles, this volume), textbook corpora (Chujo et al., this volume), translation and comparable corpora (cf. the various chapters in Section B of this volume), and so on. Corpus texts may be selected, graded or tagged for level (Huang & Liou 2007), and even simplified language in the form of graded readers may be used to create a corpus at an appropriate level of difficulty for a given population (Allan 2009; Cobb 2014). These are all hugely valuable initiatives, but many such corpora are not freely available outside the classroom or off campus, especially for long-term use after the end of a course: for DDL to continue, the learner needs stable access at any time and in any place.

Similarly, attempts have been made to produce concordancers and other tools which are more user-friendly, accessible and relevant to language learners (see Kaszubski 2006). Though several studies (e.g. Yoon & Jo 2014) find that modern concordancers are in fact less difficult to use than previously feared, perceived difficulties and latent technophobia can nonetheless be real obstacles (Lam 2000). AntConc (Anthony 2011) is among the best-known free concordancers, and though it needs downloading, it comes with a simple interface to just the essential tools rather than attempting to include all the bells and whistles which researchers have come to appreciate. Cobb (2014) also offers a range of tools on his Compleat Lexical Tutor website embedded into other CALL packages (e.g. clicking on a word in a text can open a pop-up with the relevant concordance), which may also go some way towards reducing the burden of learning how to use the concordancer. Again, solutions such as these are certainly useful, but do still require the mastery of new types of tools.

A third approach might be to abandon the more radical aspects of DDL: rather than presenting it as “revolutionary” (Johns 1990: 14), it might help to build bridges with what learners already know and do (cf. Tyne 2012) and thus to reframe it as potentially “ordinary” practice (Boulton & Tyne 2014). As Gavioli (2009: 44) maintains, “none of the activities described [by Johns] is new to the English teaching setting”. For example, it is common practice to ask learners to observe meaning and use in context, to work with authentic texts, to look at multiple examples of the same target item and so on, even though in conventional approaches this may most typically be done on the board or on paper. Similarly, there may be situations in which learners already explore large quantities of language on computer. This occurs when they use regular Internet search engines to find answers to their language questions. In other words, when they spontaneously use the web as a ‘corpus’ and the search engine as a ‘concordancer’.

Johns viewed DDL as “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” (1988: 13). Society, education and technology have all evolved considerably since, and one can only speculate as to what Johns would have been doing had he had access to the tools and content of the web as we know it today. It would surely be ironic if DDL researchers and practitioners were to stick dogmatically to academic, technological and educational dictates that are 30 years old while ignoring new developments and new ways of exploiting them.

This section has argued that Johns’ DDL is not inextricably linked to the traditional notions of corpus and concordancer, and there seems even less reason for this to be the case in a wider view today. If, then, it is possible to show that the world web can provide pedagogically relevant data, and that search engines such as Google allow learners to engage with this data in similar ways to DDL, then there seems to be little reason to automatically exclude them from the general DDL paradigm. Rather than forcing learners to adapt to corpus linguistics, this may allow us to adapt corpus linguistics to learners’ everyday practice.

3. Web as corpus

The objective of this section is to see if the web shares some of the features of a conventional corpus, discussing the failings often attributed to it from a corpus-linguistic perspective and relating that to pedagogical requirements. The status of the ‘web-as-corpus’ has aroused considerable debate. Several books (e.g. Hundt et al. 2007) and journal special issues (e.g. Kilgarriff & Grefenstette 2003) have been devoted to the question, and there was even an ICAME debate in 2011 entitled “Do we still need language corpora?” Passions are roused on both sides, but pedagogical criteria must be paramount here. The textbook definition of a corpus given earlier (a large collection of authentic texts in electronic format, designed to be representative of a language variety) is not always as straightforward as it might sound even within the field of corpus linguistics, where there are “several criteria that, if met, define a prototypical corpus, but the criteria are neither all necessary nor jointly sufficient” (Gilquin & Gries 2009: 6).

Clearly, the web fails the textbook definition on several counts, but then so would many other corpora. It is certainly not carefully balanced, but the very concept of representativeness is controversial and most corpora can be criticised to an extent on this charge (Kilgarriff & Grefenstette 2003: 333). On the other hand, given that “the web itself... [is a] huge source of language that is available in the classroom or the study at home” (Sinclair 2004: 297), and that learners are spontaneously using web sources for their learning or other personal interests or professional needs (e.g. Sockett & Toffoli 2012), the question of representativeness loses some of its impact. No corpus is neutral, and any two corpora will be non-neutral in different ways. Second, the size and composition of the web are unknown (e.g. Lüdeling et al. 2007) and probably unknowable in any definitive way, but the end-user may not be aware of what texts exactly are in, say, on-line versions of many linguistic corpora, and can only take the compilers’ word for it. Third, individual searches are not entirely replicable as the web fluctuates over time (e.g. Wu et al. 2009), but this can be seen as an advantage representing the state of the language (cf. Volk 2002); besides, the same is also true of monitor corpora, including the influential Bank of English. Fourth, “trust

the text” (Sinclair 2004) can be a hugely liberating leitmotiv, but its corollary of ‘garbage in/garbage out’ may be equally valid. Query results should always be interpreted critically rather than accepted at face value, whether the web or the BNC (Burnard 2002). Fifth, the web is not PoS-tagged or lemmatised, but nor are many other corpora; and while this limits some types of research, reducing the options available can simplify life for non-specialist users. Finally, and perhaps most unsettlingly, the web is extremely ‘noisy’ with its endless reduplications, spam, lists, nonsense pages, and so on, with innumerable different types of texts from widely varying authors writing for very different purposes all mixed up. But the same can be said (if to a lesser degree) of many semi-automated corpora compiled from the web, from COCA (Davies 2009) to the billion-word WaCKy corpora (Baroni & Bernardini 2006). In any case, this is arguably all part of “the mush of general goings-on” of real language in use (Firth 1957: 187).

None of these objections stop linguists using the web as a ‘quick and dirty’ source of language data for *ad hoc*, everyday concerns, as witnessed by the series of “breakfast experiments” by Mark Liberman on *LanguageLog* (<<http://languagelog.lidc.upenn.edu>>). It is indeed a “fabulous linguists’ playground” (Kilgarriff & Grefenstette 2003: 345). Similarly, Hoey (2012) makes no apologies about using a Google search for *flood* to “show that Macmillan [dictionaries] was getting it right.” Crystal (2011) devotes an entire book to *Internet Linguistics*, and in traditional linguistic research, the web increasingly serves as a useful point of comparison (e.g. Joseph 2004; Rohdenburg 2007; Rosenbach 2007) – and for good reason. Empirical linguistic research tends to assume that “language is never, ever, ever random” (the title of Kilgarriff’s 2005 paper), and even with all its noise and other problems, the sheer size of the web means that ‘correct’ forms can often be identified easily where they are several orders of magnitude larger than deviant forms (Kilgarriff & Grefenstette 2003: 342). In support of this, web searches often give results that are close to traditional corpora (e.g. Rohdenburg 2007; Mondorf 2007), and even to native-speaker judgements (Keller & Lapata 2003). Sinclair himself was a proponent of the axiom that there is “no data like more data” (2001: ix), a motto also taken up by Google researchers (Franz & Brants 2006). Figures vary, but Google’s current estimate is that it covers over 60 trillion web pages.³

The only real conclusion is that the issue of whether the web is (or can be used as) a corpus is a personal one. For Sinclair (2005: 21), it is simply a fact that “the World Wide Web is not a corpus”, while many such as Adolphs (2006: 33) are now prepared to admit that “we would not *normally* refer to the web as a corpus...” (emphasis added). The alternative position seems to be gaining an increasing number of advocates: for Kilgarriff and Grefenstette (2003: 334), “the answer to the question ‘Is the web a corpus?’ is yes”; for McCarthy (2008: 566), “the Internet is simply a huge corpus”; for Crystal (2011: 10), “there has never been a language corpus as large as this”; for Rundell (2000: n.p.), we can “think of the Internet itself as a giant corpus, and... use a standard search engine to find instances of a word or phrase”; and back to Kilgarriff (2001: 473), “the corpus of the new millennium is the web”. It is not just a question of research cultures: Lou Burnard was closely involved in the creation of the BNC, one of the most carefully compiled large corpora to date, yet affirms that “corpora have had their day: Google is the future” (personal communication). The ‘web-as-corpus’ is

³ <<https://www.google.com/insidesearch/howsearchworks/thestory>> (20 March, 2014).

not perfect, but nor are any other corpora; indeed, as Sinclair (2005: 98) was the first to admit, “it is important to avoid perfectionism in corpus building. It is an inexact science.”

Ultimately, if even (some) linguists can overcome qualms about using web data for everyday concerns as well as their formal research pursuits, then it would seem unreasonable to banish it from language learning which has its own requirements: the decision should be *pedagogically* driven rather than based on *research* criteria which are of little relevance in the classroom. While the web may not be a prototypical corpus in terms of linguistic research, we can at least treat it as a “corpus surrogate” (Bernardini et al. 2006: 10) which may be fit for purpose as far as language learners are concerned. Whether the web is or is not a corpus, it contains data which can be useful for language learners.

But it may be counterproductive to consider the web merely as a second-rate corpus: its specific characteristics provide a number of advantages for language teachers and learners. The web itself is fast and flexible, free and available almost anywhere in the world; it requires no download, its data source doesn’t crash, and it doesn’t impose limits on the number of simultaneous users. It includes enormous quantities of recent data of all types – whatever they want is probably there somewhere in any of the world’s written languages. Though reliable transcriptions of everyday conversational language are relatively scarce, scripts and transcripts of talk shows, moves and sitcoms have been found to be remarkably similar to spontaneous speech (Quaglio 2009; Dose 2012; Forchini 2012); targeted web searches thus provide spoken data which is appropriate for language learning purposes. Where learners are familiar with such formats from their informal online viewing or reading habits, they can all the more easily appropriate the data. The web as a whole is also already familiar to learners, especially via search engines such as Google, and this will be crucial in locating the evidence they are looking for.

4. Google as concordancer

Even if they do not use the web directly, many corpus linguists (and others) consult corpora which are derived from web data. The BYU corpora (see e.g. Davies 2009) are a classic example, offering hundreds of millions of words via a free online interface. Individual users can compile their own corpora from the Internet, selecting specific texts or partly automating the process with a tool like BootCat (Baroni et al. 2006), then opening the files in a free concordancer such as AntConc (Anthony 2011). But the mastery of the software is again likely to be an unrealistic prospect for users who do not have specific, regular needs. The web itself can be searched via a number of ‘linguistic’ search engines, notably WebCorp (Renouf et al. 2007) and KWicFinder (Fletcher 2007). These allow comparatively sophisticated language queries and linguistically-friendly output from the Internet, but require almost as much training as regular concordancers. They are also likely to be off-putting for frequent web users, as the tremendously high speed of regular search engines drives down tolerance levels for slower tools such as these.

The most obvious software for accessing the web itself as a ‘corpus’ is a general-purpose search engine like Google. Because it is familiar to most users it arguably provides a way in to corpus work (cf. Gao 2011: 262). Support for this comes from the fact that many learners, upon their first encounter with corpora, spontaneously compare them to Google (Sun 2007).

For Littlemore and Oakey (2004: 97), it can certainly be “used as a language reference tool, duplicating functions such as... [those of a] concordancer”; and Sharoff (2006: 64), while noting that “Google is a poor concordancer”, does not question that it *is* a concordancer of sorts.

But in addition to the limitations of the web as corpus outlined above, search engines introduce new problems of their own. In particular, Google is something of a black box:⁴ we do not know what pages are indexed (though it is certainly only a portion of the entire web), and the user has little idea of how the results are retrieved or ordered, and can do little to change this except submit a new query with different parameter settings or search terms (Bergh 2005). It is not just the web that fluctuates over time, but also the search algorithms used; so identical queries will provide different results, making it difficult for a teacher to follow a student’s work – compounded by the fact that Google uses cookies to store information about previous searches, thus affecting future results. The frequency figures reported are particularly problematic as they are only estimates; so Google may attribute millions of hits to an item but only return a few hundred snippets, making it impossible to check. Google and other search engines also provide results which are simply wrong for some search types, such as returning vastly more hits for “*protect against the*” as a phrase than for “*protect against*” – a logical impossibility.⁵

It can also be difficult for learners to interpret the results, in particular in deciding how frequent is frequent ‘enough’ to indicate acceptability. For Wu et al. (2009: 253), web searches are probably “a good enough indication for language learning purposes”, which is the crucial requirement. But how are learners to make sense of the alleged millions of results for *discuss about* when their teachers tell them it is wrong?⁶ While learners may appreciate the more rigid structure of a corpus after exploring the web, they may conversely come to think of a conventional corpus as no more than a smaller (and hence less desirable) version of the web (Conroy 2010).

Search engines are primarily designed for retrieving content, and are thus inevitably less helpful for language analysis than concordancers and other software specifically designed for this purpose. They do however allow a variety of search types, notably “*hunting*” and “*browsing*” (Hawkins 1996), which are essentially the same as concordance searches for specific information and serendipitous foraging respectively. Google does not allow explicitly linguistic search syntax, and ways round its limitations can be time-consuming and still very approximate (Sha 2010). But the reduced number of options available can in some respects

⁴ Though see Google’s Search Quality Rating Guidelines (Version 1.0): <<https://static.googleusercontent.com/media/www.google.com/en//insidesearch/howsearchworks/assets/searchqualityevaluatorguidelines.pdf>>.

⁵ Many such anomalies are reported in the blog postings of the late Jean Véronis, e.g. 5 Billion ‘the’ have Disappeared Overnight; Yahoo’s Missing Pages; Crazy Duplicates; Google: Mystery Index, and many more. Yet we are still left with Google: The Largest Linguistic Corpus of all Time. (Technologies du Langage: Actualités, Commentaires, Reflexions: <<http://blog.veronis.fr>>.)

⁶ My search at the time of writing (March 2014) gives 27.2 million Google hits for “*discuss about the*” vs 190 million for “*discuss the*”. It may also be that *discuss about* is a developing norm in English as a Lingua Franca; Davies’ GloWbE corpus (<<http://corpus2.byu.edu/glowbe>>) suggests it is particularly common in Indian and Bangladeshi English.

make life simpler for non-specialist users, as Johns found with his own concordancer: “the more complex the program, the more inaccessible it may become” (1986: 156). Though the output snippets are not entirely dissimilar to concordances, the presentation is not linguistically ideal. And of course, Googling simply does not feel like a ‘serious’ pursuit; but as we have seen, if linguists can use it at least informally for this purpose (Bernardini et al. 2006: 37), then *a fortiori* language learners whose requirements are different. It is perhaps a truism that “Googleology is bad science” (Kilgarriff 2007), but language learners are not engaged in ‘science’ in the same way as corpus linguists are, even in the learner-as-researcher paradigm – again, pedagogical considerations take precedence. In any case, no concordancer is ideal (Kaszubski 2006), and though general-purpose search engines may be the least ideal of all, there seems to be nothing to stop us treating “Google as a quick ‘n’ dirty corpus tool” (Robb 2003). Optimistically, it may even be that the messiness of web data and limitations of search engines will foster language awareness and critical thinking about language (Milton 2006). We can accept the imperfection of the web and its search engines and still conclude that web data are useful for pedagogical purposes (e.g. Sha 2010: 389).

As McCarthy (2008: 566) has it, “we are, all of us, corpus users, because we use the Internet”, and the familiarity itself constitutes a tremendous advantage of Google over dedicated concordancers (Park 2012; Smith 2011: 297). Google can be used for language learning *per se*, but perhaps more importantly still as a reference resource to find suggestions and check one’s own writing or translation (e.g. Geluso 2013). For example, the MeLLANGE project (Kübler 2011) found over 93% of professional translators using Google, and barely 20% using corpora. Teachers and learners use search engines in this way for their own language purposes without ever having heard of corpora or data-driven learning (cf. Conroy 2010; Geluso 2013), just as dictionary users are not expected to be lexicographers. Dictionaries are easy to use in a rudimentary fashion even on first encounter, though only training will help learners to make the most of them. Imperfect Google use should not imply that learners should be discouraged from using Google at all, any more than imperfect dictionary use implies that dictionaries should be avoided until they are fully mastered (cf. Nesi 2000). Using the web for language searches is useful in itself, and training in more effective Google use such as inverted commas for fixed phrases is likely to be “one of the best tips to teach your students” according to Dudeney (2000: 22), not least because it should be very quick compared to training in using a concordancer. This may be enough for some learners, while others, especially those whose language needs are likely to continue for some time, may progress to prototypical corpus use. Beginning with search engines should provide a “seamless and unnoticeable” transition to concordancer use (Bernardini et al. 2006: 37) where this is appropriate; from there, learners can move on directly to corpora or pass through any number of half-way stages, from the search engines (often powered by Google) on newspaper and other websites to the student favourite Linguee (<<http://www.linguee.com>>), which allows parallel searches of translated texts (see Buyse & Verlinde 2013).

5. Searching the web

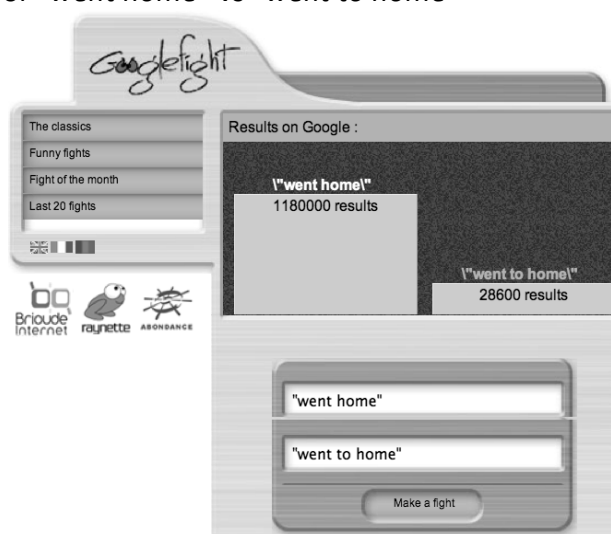
Tools with very specific uses are likely to be opened only occasionally, and with limited use comes limited familiarity and ease, thus engendering a vicious downward spiral culminating in no use at all (cf. Boulton 2011b). There is a greater chance that students will continue

linguistic queries with a tool that is already part of their daily lives, so any work in this direction is less likely to be lost. And because such linguistic uses of Google build on existing skills, they are by nature transversal and transferrable, so any training is likely to feed back into real-world Internet use, fostering digital literacy in a much broader sense: “an accomplished corpus user also becomes a more efficient Googler” (Philip 2011: 65).

This section is not intended as a how-to guide, since Google itself and other websites provide many tips on use, but attempts to highlight some of the main features that can be useful for language learners and teachers. Specific features inevitably come and go: the wonderwheel function, cache feature and timeline displays are sorely missed, but no doubt others will appear. Different versions of Google are available; it may be more relevant to use Google.fr for French, for example, or Google Scholar for academic language (preferred by some students in Sun 2007). As an aside on this last point, better results are likely to be achieved from COCA than from Google Scholar (see the debate between Brezina 2012 and Davies 2013, comparing the two), but the basic premise remains unchanged: in the short term, easy results via Google Scholar obviate some of the need for training to use COCA. This can be seen in the fact that Davies finds that Brezina substantially misrepresents the data available in COCA, which he attributes to “inexperience in knowing how to query the corpus to retrieve the desired results” – if a published scholar has difficulty querying COCA for this, students are likely to perform even less well.

Basic training in using search engines to access language information partly concerns encouraging learners (even native speakers; see Young 2011) just to think about what they are doing, ensuring they search for useful items and view the results critically. While frequency returns from search engines are notoriously unreliable, as we have seen, comparative frequencies may be useful in some cases if interpreted carefully. GoogleFights (<<http://www.googlefight.com>>) gives a graphically simple indication of normal usage when frequencies vary by orders of magnitude, and these comparisons can be used to sensitise learners to the dangers of relying on even thousands of hits as evidence. Figure 1 suggests that *went home* is hugely more frequent than *went to home*, for example.

Figure 1. GoogleFight for “*went home*” vs “*went to home*”



Google has some features which present an advantage over traditional concordancers, and not just in terms of speed and coverage (Sha 2010; Sun 2007). The interface is of course extremely user-friendly (Bernardini et al. 2006: 37), and exists in many different languages for users of different L1s. One of the basic concepts in DDL is pattern-detection; this can even work with Google images. Figure 2a shows that a *can* is typically associated with drink and to a lesser extent food, while a *tin* refers to the metal and to food but not drink. Figure 2b show that the English *nut* has far wider coverage than the French *noix* ≈ ‘walnut’ (cf. Boulton & Tyne 2014).

Figure 2a. Google images for *can* vs *tin*

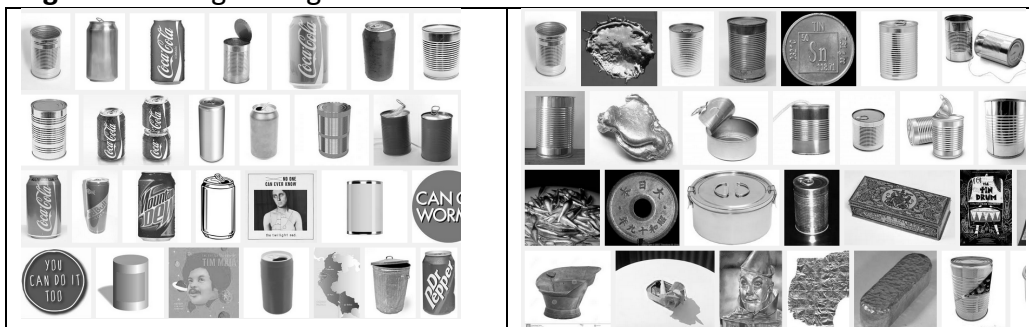
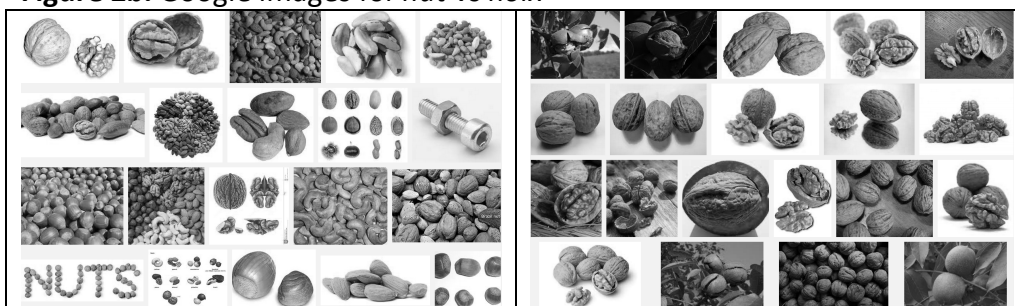
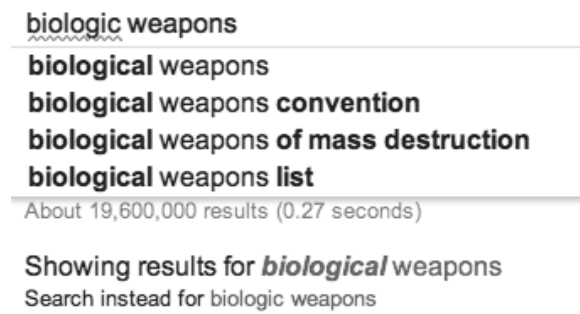


Figure 2b. Google images for *nut* vs *noix*



Google can be introduced as a regular activity with a minimum of metalanguage or corpus terminology (cf. Frankenberg-Garcia, forthcoming). Although ‘linguistic’ queries can prove difficult, its search functions incorporate a certain flexibility not available with regular concordancers which require exact chains (cf. Park 2012). At the search stage, Google underlines words not in its dictionary in the same way as most word processing packages, and proposes alternative formulations: a student who types in *biologic weapons* will be given the prompt: “showing results for *biological weapons*” (Figure 3a). Search terms are returned in different orders, which can be helpful for noun phrases: *tall mountain world* gives *tallest mountain in the world* as well as *tall mountains around the world*.

Figure 3. Google search for *biologic weapons*

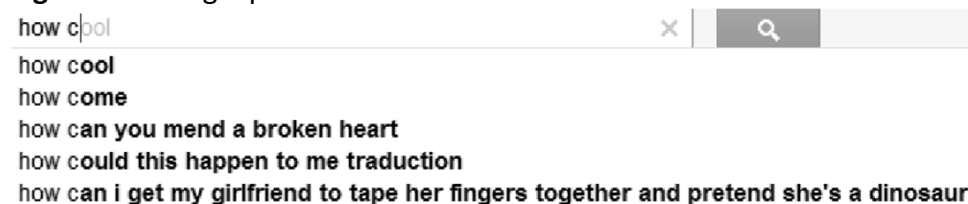


This interpretation of search functions can be expanded using the Google Instant predictive function (if switched on in 'settings') to produce all sorts of unexpected results such as those in Figure 4, obtained by beginning to type *can g...* or *how c...* respectively. This can be the cause of much mirth and give rise to genuinely spontaneous discussion in class, but it does also have a serious point in showing how questions are formed in English.

Figure 4a. Google instant search for *can g...*



Figure 4b. Google predictive search for *how c...*



Searches can be refined in the regular search window, most notably by using inverted commas to search for a phrase akin to ordinary corpus searches. For example, Google today returned 477 million hits for "*the same*" vs 12 million for "*a same*" (each time in inverted commas), suggesting that the former is more usual but that the latter might be possible on occasion. A perusal of the snippets (e.g. Figure 5) reveals contexts such as *a same-sex relationship*, *a same-sex couple* and *a same-day loan*, where *same* is used in a very specific way that can give rise to noticing and discussion in the classroom.

Figure 5. Google search for “a same”

Dallas Methodist minister suspended after complaint over his ...
thescoopblog.dallasnews.com/.../dallas-method... ▾ The Dallas Morning News ▾
18 hours ago - ... Church that he has been suspended from all clerical responsibilities
after a complaint was filed for his performance of a **same**-sex marriage.

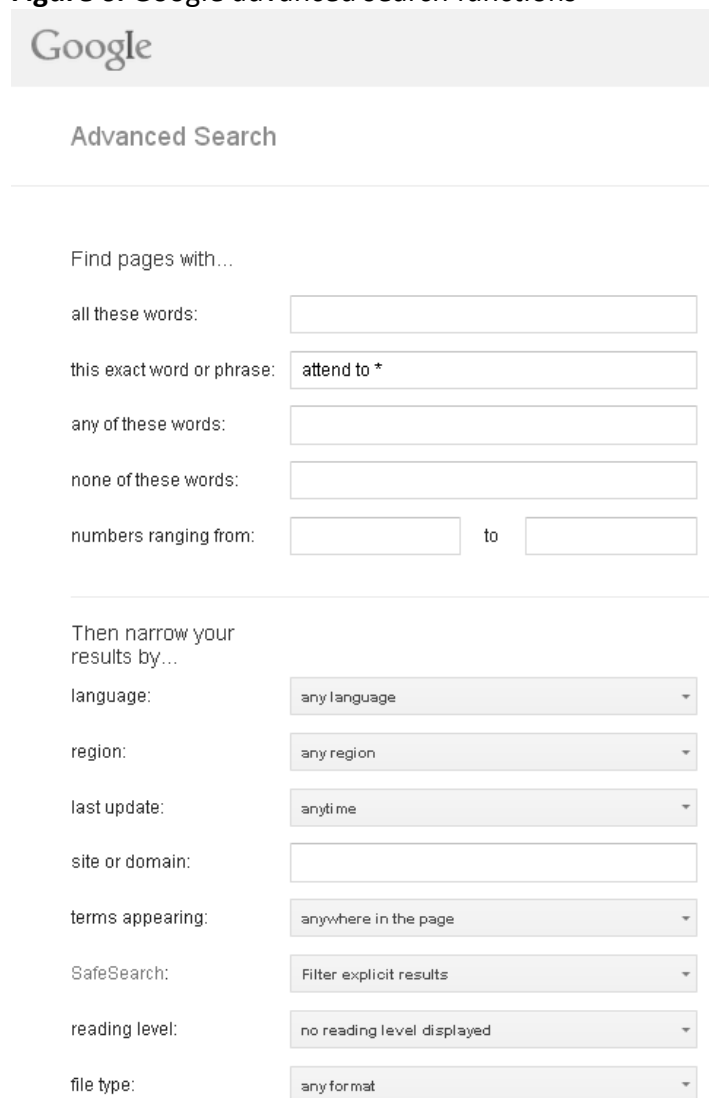
Notice regarding same sex marriage and domestic partnerships
https://www.sos.wa.gov/.../Notice-regarding-same-sex-marriage-and-do... ▾
If you are a **same**-sex couple and both of you are under 62, you will be converted to
marriage. If you filed a domestic partnership in a city, county, or other state or ...

How to Get a **Same** Day Car Loan: 5 Steps - wikiHow
www.wikihow.com > ... > Cars > Car Finances ▾ wikiHow ▾
How to Get a **Same** Day Car Loan. More than a few car buyers may want to take their
time when financing a vehicle, but others might want to find the quickest ...

In conjunction with inverted commas, an asterisk can be used as a wildcard for any individual word within a phrase. For example, students can check the type of noun found in “*attend the **” and “*attend to the **”, preposition use being a common query type in regular L2 corpus consultation (e.g. Kennedy & Miceli 2001: 83). This can be done in the regular search window or, as shown in Figure 6, in the advanced search option (by clicking on ‘settings’, or the cogwheel after a search to refine it). Among various useful features, it is possible to exclude unwanted words, which can limit irrelevant hits in the case of polysemous items, e.g. *crane –bird* to include pages with the word *crane* but not *bird*, thus focusing on the machine. Searches can also be restricted by language and date, or to a particular file type (pdf format may elicit more research articles than general web searches) or domain (e.g. .edu or .ac.uk or .univ.fr for academic websites). Bergh (2005) shows that such domain-specific searches for ‘slices’ of the web generally provide more rewarding results than indiscriminate web searches, especially for questions of frequency. Similarly, Googling within a particular site can be particularly useful for academic writing as it is possible to limit the search to a particular online journal. The ‘reading level’ feature, though imperfect, can be useful for lower-level learners as it can restrict searches (albeit imperfectly) to texts of appropriate levels of difficulty, since the complexity of authentic language has been claimed as a barrier to DDL on occasion (e.g. Allan 2009). Google advanced has to be reset every time (cookies notwithstanding), but default settings can be defined in Google CSE (Custom Search Engine). Geiller (in press; see below) describes how he installed this on his coursepage so that his students would only be searching on sites he preselected for them, mainly British and American newspapers.

If it is difficult to formulate a linguistic query, it is no less difficult to interpret the results for language purposes. The output is in the form of ‘snippets’ which, though not entirely dissimilar to concordance lines, are not aligned, and the default option shows only 10 hits per page (this can be changed under ‘settings’). Scrolling through the pages soon shows that the alleged millions are not forthcoming, and the first few pages of hits are often unreliable as site owners can manipulate the rankings of their sites, even paying to move up the scale. So rather than ending with the snippets as one might end with concordance lines, it is often advisable to open a few selected pages and locate the same item again in context (Conroy 2010) using CTRL+F. Though users may be tempted to trust in a brand such as Google, they should not forget that it is a commercial venture, and results should always be critically assessed (cf. Hargittai et al. 2010).

Figure 6. Google advanced search functions



The image shows the Google Advanced Search interface. At the top is the Google logo. Below it is the heading "Advanced Search". The interface is divided into two main sections: "Find pages with..." and "Then narrow your results by...".

Find pages with...

- all these words:
- this exact word or phrase:
- any of these words:
- none of these words:
- numbers ranging from: to

Then narrow your results by...

- language:
- region:
- last update:
- site or domain:
- terms appearing:
- SafeSearch:
- reading level:
- file type:

6. Previous studies

Arguments to the effect that learners can derive benefit from such-and-such an approach are useful, but really acquire weight when supported by empirical evidence. The objective of this section is not to provide new evidence, but to briefly present existing research examining learners' use of search engines to explore the web for language purposes. Though there are a number of passing references in various papers, few studies have so far focused specifically on this.

Scheffler (2007) begins with a survey of web uses for language learning purposes among 100 Polish EFL learners. Unsurprisingly, dictionaries and authentic texts on the web top the list of their preferred resources, but 10% of respondents used the 'web as corpus' – the term used by the researcher covering queries "to search for or to check the use of various linguistic items" (p. 141). By way of contrast, only one respondent used a conventional corpus (the BNC). Of the 20 teachers questioned, six (i.e. 30%) used the web in this way for personal use, and two the BNC. Scheffler goes on to provide suggestions for using general search engines

for various types of linguistic query based on the traditional corpus query model. These are inspired by errors taken from learners' writing, so the author knows in advance what outcome to expect; though learners may find it difficult to initiate such searches without being told what points to look for, bringing their attention to errors and then asking them to consult a corpus is standard fare in the DDL literature (e.g. Chambers & O'Sullivan 2004; O'Sullivan & Chambers 2006).

In an experiment on writing a short story, Philip (2011) divided her students *post hoc* into small groups depending on their preferred resources, classifying them as corpus users (six students), dictionary users (three), Googlers (three), and mixed-resource users (twenty). Though the corpus was found to be effective, this was partly because it was used by the "archetypal 'good language learner'" (p. 66), i.e. students who were motivated to discover and experiment with a new tool to add to their arsenal. Google was less effective partly for the opposite reasons, encouraging elementary single-word searches to check frequencies or identify possible expressions in the snippets (or occasionally by opening the full page). The aim of the study was to test the advantage of corpus use over simple Googling, and this certainly seems to have been borne out: "the Internet, while undeniably useful as a linguistic ready-reckoner, fosters neither accuracy nor variety in the acquisition and use of lexis and phraseology" (p. 66). Where corpus training and use is possible, this is certainly to be preferred, and an alternative outcome would have been surprising to say the least.

Other studies which focus on web uses tend toward more positive results. Todd (2001), for example, asked postgraduate Thai students in science and engineering to provide 10 examples derived from the now defunct AllTheWeb search engine which they used to help correct lexico-grammatical errors indicated in their writing. No discussion is provided of the status of the data, the researcher seeming to assume that they can be considered concordances. The patterns induced from these 'concordances' generally matched those found in reference books, and 18 of the 23 participants were able to correct their errors successfully. Detailed analysis shows that these intermediate-level learners were able to locate relevant information from the web, induce patterns and apply them to their own writing, even with very minimal training.

The use of inverted commas in Google to search for exact phrases was key in a study by Acar et al. (2011). Twenty minutes was spent showing eight Japanese engineering students how and why to use the feature to check for possible errors in their own writing. They chose 4-word chunks which they had doubts about and encased them in inverted commas to check for frequency on the web. The general proposal was that fewer than 100 hits should be taken as a sign to modify the text (especially for prepositions, articles, etc.) and then perform a new search with the revised chunk. Though most of the revised sentences still contained errors, an average 24% showed improvement in clarity or grammatical accuracy; the least successful student improved 16% of sentences, while the most successful improved 31% (which still, of course, leaves a majority of sentences which were not improved). Shei (2008a, 2008b) proposes a similar method, first checking the frequency of a target word, then the frequency of that word plus the word immediately to its right, and so on (e.g. *found*, *found to*, *found to be*, *found to be infected*, etc.). There comes a point when the next word leads to a dramatic decrease in frequency, which is where learners may be induced to

check their formulations and try alternative possibilities. Though Shei does not provide classroom evidence, the basic approach in both these studies leaves open the possibility that such individual searches may some day be automated, dividing a document into segments of a specified length (text-tiling) and highlighting those segments with low frequency which may therefore deserve attention.

One of the authors of the Acar et al. paper extended their line of enquiry to formulaic sequences. In this more rigorous study, Geluso's (2013) Japanese students were briefly introduced to Google's inverted commas as a way to find normal usage based on frequency; they then selected segments they were unsure of in their own writing, and were asked to use Google to help revise them. Four native-speaker raters compared 334 phrases before and after revision, half based on web searches and half without. Each of the raters independently scored the Google-informed revisions more highly, a significant result overall. Geluso's conclusion is that:

while the web and Google are not designed to be corpus and concordancer, respectively, they can be defined as such given their characteristics and functionality. Training students to perform double quotation mark searches on Google is a relatively simple matter, and considering the ubiquitous nature of the search engine, many students may already be engaging in such behaviour. (p. 155)

Rather than take the web as a whole, Geiller (in press) personalises Google's CSE function to search 28 British and American websites for his relatively advanced learners, encouraging them to discover new linguistic contexts for particular items. For example, a search for *ban assault weapons* brought up complex noun phrases and many other uses which the students could reuse in their own writing (e.g. *introduce new legislation to ban assault weapons* as well as *an assault weapons ban*, and so on). Careful use of inverted commas and asterisks in particular allows work on collocations: "*play a * part in*" brings up not only *big* and *large*, but also *major*, *significant*, etc. Geiller argues that this provides learners with large chunks of raw material for subsequent analysis and integration to their language knowledge. Evidence comes in the form of 129 errors derived from students' essays, all of which were deemed 'untreatable' in the sense that the learners' current stage of knowledge would not allow them to correct the errors without other input (see Ferris & Roberts 2001). Use of the customised search engine enabled the learners to correct 52% of them appropriately, as against 28% inappropriately.

Park is particularly interested in the *processes* involved rather than learning outcomes alone. A 2012 paper focuses on use of Google CSE by three Chinese business students to access 50 selected academic papers online, thus coming closer to traditional corpus use inasmuch as the exact contents are known. Search logs show that they used the tool with varying frequency, and that on average 58% of 'transactions' (i.e. query sequences) featured multiple searches. Overall, 43% of the transactions did not lead to changes, but 53% led to improved writing, compared to only 4% of changes for the worse. A more in-depth case study (Park & Kinginger 2010) shows that the processes involved in the transactions corresponded remarkably closely to corpus searches: each involved a perceived problem which provided a hypothesis and was formulated as a search query; the results were then

evaluated and could feed back to refine the hypothesis or query until a satisfactory answer to the problem was arrived at.

The importance of tracking logs is highlighted in Pérez-Paredes et al. (2011). The focus here was on guided and non-guided use of the BNC, but results include learners' use of other online resources, notably Google to search for complex items. The experimental (guided consultation) group used Google much vastly than the control group (82% vs 7%), visiting it more often even than the recommended dictionary website. The results thus show that training in corpus use does have a direct feedback effect on other ICT uses, at least in the context of this short experiment.

Conroy (2010) noted that many of his students in Australia were already using the Internet for a variety of language learning purposes, encouraging him to develop a course introducing corpora alongside advanced web search techniques. Detailed analysis is confined to error correction among three of the students who had been trained to use concordancers and Google for language purposes. These students used a concordancer to correct only one of the 45 errors indicated, and this incorrectly; Google was used in 22 cases, of which 15 were successful. Following training, 53% claimed they would continue to use Google for language learning, ahead of corpora and concordancing at 36%, with some students explicitly saying they found searching the web more useful as more likely to provide relevant data. This is no doubt because the participants had a head start with Google, which is therefore presented as one entry point to more advanced corpora work for those willing to make the investment. The study also provides some evidence that the corpora training improved these students' web searches.

Similarly, following a successful introduction to corpora use for EFL writing, Chang (2010) also notes that a number of open comments in the post-course questionnaire referred to previous uses of Google and Yahoo for language learning purposes, though they generally cite the advantages of corpora and concordancers which are received more favourably. Opinions and findings are mixed overall, however. On the one hand, Hafner and Candlin (2007) found their learners using a legal corpora mainly as a source of information about the law (equivalent to Google searches, which they also used) rather than about language *per se*, and slowly abandoning corpora searches during the course. On the other, Park (2012) found participants focusing only on language and rarely on information content, structural organisation, etc., and expecting to continue using corpora in the future.

The small number of studies surveyed here shows that the use of web search engines directly for language learning and teaching purposes is as yet largely uncharted territory. Many of the uses described correspond broadly to DDL, although most of the work to date focuses on the web as a resource for writing (including revising and error-correction) rather than as a learning aid as such. Though there is no guarantee that corpora consultation will lead to learning in such cases, encounters with the language are likely to contribute to learning just as dictionary look-ups lead to incidental learning (see Laufer & Hulstijn 2001). More research is certainly needed, especially for longitudinal development (cf. Flowerdew, this volume). However, the interest is such that it has already given rise to new terms such

as GALL for “Google-assisted language learning” (Chinnery 2008), and even, modelled on DDL, “Google-driven language learning” (Sha 2010).

7. Conclusion

This paper has argued that, for pedagogical purposes: (a) the web can be equated with a vast corpus insofar as it represents data that can be useful for L2 learners and users; (b) web search engines can be equated with concordancers in that they allow the user to search that data; and (c) the two together can be used in ways compatible with a data-driven learning approach. While it seems clear that none of this is prototypical DDL, the primary criteria are pedagogical, and the web+corpus can still promote many of the same advantages inherent in DDL. Given that the Internet has been around for some time now, there is surprisingly little direct empirical research to date; what there is tends to be encouraging, but is by no means overwhelming. The aim here is not to suggest that web searches are better than prototypical corpus- and concordancer-based DDL – far from it. Rather, they can be seen as one instantiation of it which may have some uses for some learners in some contexts.

It would seem to be self-defeating to rigorously apply corpus linguistic criteria to pedagogical situations, or to stick dogmatically to existing forms of DDL. Rather than picking fault, teachers and researchers should be happy that some newspaper sites provide frequency information and collocates in their results, that Amazon allows ‘search inside’ functions to provide concordance-like results, that word processing packages show search hits in a separate window with the search string highlighted, that word clouds are found in the unlikeliest of places well outside academia, and in particular that search engines such as Google allow L2 users to explore the target language on their own. It seems likely that many learners around the world are already searching the web in ways not entirely dissimilar to DDL, a practice which may be actively encouraged by their teachers while remaining invisible in the DDL research literature. The approach is in many ways attractive, offering as it does a familiar and easy way to begin simple DDL and which can bring immediate benefits. To reach a wider audience – the “corpus mission” as Römer (2010) would have it – there is also something to be said for encouraging the perception of DDL as *ordinary* practice by building on common behaviours both inside and outside the classroom; it is only when such uses become ‘normalised’ (Pérez-Paredes et al. 2011: 247) that they can be considered completely integrated and can reach their full potential.

The main conclusion is pragmatic and practical rather than dogmatic or ideological: if an approach or technique is of benefit to the learners and teachers concerned, it should not be ruled out automatically. As so often, there is likely to be a payoff between how much the teachers/learners are prepared to put in (ideally as little as possible) and how much they want to get out (ideally as much as possible), i.e. as they seek maximum return on investment. The optimum will be at some variable point in between, or more likely a movement along the continuum – gradually investing more and more, until such time as the extra benefits do not justify the extra costs. Though principled corpora and dedicated tools certainly add value to the approach, the question in any specific case is whether such value is worth the investment here and now. Such a cost-benefit analysis will produce different results for different individuals and groups with different needs and preferences, facilities and constraints. Though learners may not be using search engines very well, they are already

getting results, and a little further training is likely to increase their efficiency. For some L2 learners this will be enough, especially where language needs, motivations and ambitions are modest, or where time and resources are limited. For others, especially at more advanced levels and with specific needs and sustained motivation, it may provide a way in to more advanced uses of corpora and concordancers.

References

- Acar, A., Geluso, J. & Shiki, T. 2011. How can search engines improve your writing? *CALL-EJ* 12(1): 1–10.
- Adolphs, S. 2006. *Introducing Electronic Text Analysis: A Practical Guide for Language and Literary Studies*. London: Routledge.
- Allan, R. 2009. Can a graded reader corpus provide 'authentic' input? *ELT Journal* 63(1): 23–32.
- Anthony, L. 2011. *AntConc*, version 3. Tokyo: Waseda University. <<http://www.antlab.sci.waseda.ac.jp>> (17 February, 2013).
- Aston, G. 1997. Small and large corpora in language learning. In *Practical Applications in Language Corpora*, B. Lewandowska-Tomaszczyk & J. Melia (eds), 51–62. Lodz: Lodz University Press.
- Baroni, M. & Bernardini, S. (eds). 2006. *Wacky! Working Papers on the Web as Corpus*. Bologna: Gedit.
- Bergh, G. 2005. Min(d)ing English language data on the web: What can Google tell us? *ICAME Journal* 29: 25–46.
- Bernardini, S., Baroni, M. & Evert, S. 2006. A WaCky introduction. In *Wacky! Working Papers on the Web as Corpus*, M. Baroni & S. Bernardini (eds), 9–40. Bologna: Gedit.
- Biber, D., Conrad, S. & Reppen, R. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Boulton, A. 2010a. Data-driven learning: Taking the computer out of the equation. *Language Learning* 60(3): 534–572.
- Boulton, A. 2010b. Data-driven learning: On paper, in practice. In *Corpus Linguistics in Language Teaching*, T. Harris & M. Moreno Jaén (eds), 17–52. Bern: Peter Lang.
- Boulton, A. 2011a. Data-driven learning: The perpetual enigma. In *Explorations across Languages and Corpora*, S. Gońdz-Roszkowski (ed.), 563–580. Frankfurt: Peter Lang.
- Boulton, A. 2011b. Bringing corpora to the masses: Free and easy tools for interdisciplinary language studies. In *Corpora, Language, Teaching, and Resources: From Theory to Practice*, N. Kübler (ed.), 69–96. Bern: Peter Lang.
- Boulton, A. 2012. Hands-on/hands-off: Alternative approaches to data-driven learning. In *Input, Process and Product: Developments in Teaching and Language Corpora*, J. Thomas & A. Boulton (eds), 152–168. Brno: Masaryk University Press.
- Boulton, A. & Tyne, H. 2014. *Des Documents Authentiques aux Corpus: Démarches pour l'Apprentissage des Langues*. Paris: Didier.
- Braun, S. 2005. From pedagogically relevant corpora to authentic language learning contents. *ReCALL* 17(1): 47–64.
- Braun, S. 2010. Getting past 'groundhog day': Spoken multimedia corpora for student-centred corpus exploration. In *Corpus Linguistics in Language Teaching*, T. Harris & M. Moreno Jaén (eds), 75–97. Bern: Peter Lang.

- Brezina, V. 2012. Use of Google Scholar in corpus-driven EAP research. *Journal of English for Academic Purposes* 11(4): 319–331.
- Burnard, L. 2002. Where did we go wrong? A retrospective look at the British National Corpus. In *Teaching and Learning by Doing Corpus Analysis*, B. Kettemann & G. Marko (eds), 51–70. Amsterdam: Rodopi.
- Buyse, K. & Verlinde, S. 2013. Possible effects of free on line data driven lexicographic instruments on foreign language learning: The case of Linguee and the Interactive Language Toolbox. *Procedia: Social and Behavioral Sciences*, 95: 507–512.
- Chambers, A. & O’Sullivan, Í. 2004. Corpus consultation and advanced learners’ writing skills in French. *ReCALL* 16(1): 158–172.
- Chang, J.-Y. 2010. Postsecondary EFL students’ evaluations of corpora with regard to English writing. *SNU Journal of Education Research* 19: 57–85. <http://space.snu.ac.kr/bitstream/10371/72997/1/vol19_3.pdf> (11 April, 2011).
- Cheng, W. 2011. *Exploring Corpus Linguistics: Language in Action*. London: Routledge.
- Chinnery, G. 2008. You’ve got some GALL: Google-assisted language learning. *Language Learning & Technology* 12(1): 3–11.
- Cobb, T. 2014. A resource wish-list for data-driven learning in French. In *Ecological and Data-Driven Perspectives in French Language Studies*, H. Tyne, V. André, A. Boulton, C. Benzitoun & Y. Greub (eds), 257–292. Newcastle: Cambridge Scholars.
- Conroy, M. 2010. Internet tools for language learning: University students taking control of their writing. *Australasian Journal of Educational Technology* 26(6): 861–882.
- Cotos, E. 2014. Enhancing writing pedagogy with learner corpus data. *ReCALL* 26(2): 202–224.
- Crystal, D. 2011. *Internet Linguistics*. Abingdon: Routledge.
- Davies, M. 2009. The 385+ million word Corpus of Contemporary American English (1990-2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics* 14(2): 159–188.
- Davies, M. 2013. Google Scholar and COCA-Academic: Two very different approaches to examining academic English. *Journal of English for Academic Purposes* 12: 155–165.
- Dose, S. 2012. Scripted speech in the EFL classroom: The Corpus of American Television Series for teaching spoken English. In *Input, Process and Product: Developments in Teaching and Language Corpora*, J. Thomas & A. Boulton (eds), 103–121. Brno: Masaryk University Press.
- Dudeny, G. 2000. *The Internet and the Language Classroom*. Cambridge: Cambridge University Press.
- Ferris, D. & Roberts, B. 2001. Error feedback in L2 writing classes: How explicit does it need to be? *Journal of Second Language Writing* 10: 161–184.
- Firth, J. 1957. *Papers in Linguistics 1934-1951*. Oxford: Oxford University Press.
- Fletcher, W. 2007. Concordancing the web: Promise and problems, tools and techniques. In *Corpus Linguistics and the Web*, M. Hundt, N. Nesselhauf & C. Biewer (eds), 25–45. Amsterdam: Rodopi.
- Forchini, P. 2012. *Movie Language Revisited: Evidence from Multi-Dimensional Analysis and Corpora*. Frankfurt: Peter Lang.
- Frankenberg-Garcia, A. In press. How language learners can benefit from corpora, or not. *Recherches en Didactique des Langues et des Cultures*, 10(2).

- Franz, A. & Brants, T. 2006. All our n-gram are belong to you. *Google Machine Translation Team Research Blog*. <<http://googleresearch.blogspot.fr/2006/08/all-our-n-gram-are-belong-to-you.html>> (6 June, 2012).
- Gao, Z.-M. 2011. Exploring the effects and use of a Chinese-English parallel concordancer. *Computer Assisted Language Learning* 24(3): 255–275.
- Gavioli, L. 2009. Corpus analysis and the achievement of learner autonomy in interaction. In *Using Corpora to Learn about Language and Discourse*, L. Lombardo (ed.), 39–71. Bern: Peter Lang.
- Geiller, L. In press. How EFL students can use Google to correct ‘untreatable’ written errors. *Eurocall Review* 22(2).
- 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.
- Ghadessy, M., Henry, A. & Roseberry, R. (eds). 2001. *Small Corpus Studies and ELT: Theory and Practice* [Studies in Corpus Linguistics 5]. Amsterdam: John Benjamins.
- Gilquin, G. & Granger, S. 2010. How can data-driven learning be used in language teaching? In *The Routledge Handbook of Corpus Linguistics*, A. O’Keeffe & M. McCarthy (eds), 359–370. London: Routledge.
- Gilquin, G. & Gries, S. 2009. Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5(1): 1–26.
- Hafner, C. & Candlin, C. 2007. Corpus tools as an affordance to learning in professional legal education. *Journal of English for Academic Purposes* 6(4): 303–318.
- Hargittai, E., Fullerton, L., Menchen-Trevino, E. & Thomas, K. 2010. Trust on the web: How young adults judge the credibility of online content. *International Journal of Communication* 4: 468–494.
- Hawkins, D. 1996. Hunting, grazing, browsing: A model for online information retrieval. *ONLINE* 20: n.p. <<http://www.onlinemag.net/JanOL/hawkins.html>> (17 July, 2006, via <<http://web.archive.org>>).
- Hoey, M. 2012. Lexical priming: The odd case of a psycholinguistic theory that generates corpus-linguistic hypotheses for both English and Chinese. Paper given at *Corpus Technologies and Applied Linguistics*. Suzhou: Xi’an Jiaotong Liverpool University, 28-30 June.
- Huang, H.-T. & Liou, H.-C. 2007. Vocabulary learning in an automated graded reading program. *Language Learning & Technology* 11(3): 64–82.
- Hundt, M., Nesselhauf, N. & Biewer, C. (eds). 2007. *Corpus Linguistics and the Web*. Amsterdam: Rodopi.
- Johns, T. 1986. Micro-Concord: A language learner’s research tool. *System* 14(2): 151–162.
- Johns, T. 1988. Whence and whither classroom concordancing? In *Computer Applications in Language Learning*, P. Bongaerts, P. de Haan, S. Lobbe & H. Wekker (eds), 9–27. Dordrecht: Foris.
- Johns, T. 1990. From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *CALL Austria* 10: 14–34.
- Johns, T. 1991. Should you be persuaded: Two examples of data-driven learning. In *Classroom Concordancing*, T. Johns & P. King (eds), *English Language Research Journal* 4: 1–16.
- Johns, T. 1993. Data-driven learning: An update. *TELL&CALL* 2: 4–10.

- Johns, T. 1997. Contexts: The background, development and trialling of a concordance-based CALL program. In *Teaching and Language Corpora*, A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (eds), 100–115. Harlow: Addison Wesley Longman.
- Johns, T. & King, P. (eds). 1991. *Classroom Concordancing*. *English Language Research Journal* 4.
- Johns, T., Lee, H.-C. & Wang, L. 2008. Integrating corpus-based CALL programs in teaching English through children's literature. *Computer Assisted Language Learning* 21(5): 483–506.
- Joseph, B. 2004. The editor's department: On change in *Language* and change in language. *Language* 80(3): 381–383.
- Kaszubski, P. 2006. Web-based concordancing and ESAP writing. *Poznań Studies in Contemporary Linguistics* 41: 161–193.
- Keller, F. & Lapata, M. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics* 29(3): 459–484.
- Kennedy, C. & Miceli, T. 2001. An evaluation of intermediate students' approaches to corpus investigation. *Language Learning & Technology* 5(3): 77–90.
- Kilgarriff, A. 2001. Web as corpus. In *Corpus Linguistics: Readings in a Widening Discipline*, G. Sampson & D. McCarthy (eds), 471–473. London: Continuum.
- Kilgarriff, A. 2005. Language is never, ever, ever random. *Corpus Linguistics and Linguistic Theory* 1(2): 263–275.
- Kilgarriff, A. 2007. Googleology is bad science. *Computational Linguistics* 33(1): 147–151.
- Kilgarriff, A. & Grefenstette, G. (eds). 2003. *Web as Corpus*. *Computational Linguistics* 29(3).
- Kübler, N. 2011. Working with corpora for translation teaching in a French-speaking setting. In *New Trends in Corpora and Language Learning*, A. Frankenberg-Garcia, L. Flowerdew & G. Aston (eds), 62–80. London: Continuum.
- Lam, Y. 2000. Technophilia vs technophobia: A preliminary look at why second-language teachers do or do not use technology in their classrooms. *Canadian Modern Language Review* 56(3): 390–420.
- Laufer, B. & Hulstijn, J. 2001. Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics* 22(1): 1–26.
- Leech, G. 1997. Teaching and language corpora: A convergence. In *Teaching and Language Corpora*, A. Wichmann, S. Fligelstone, A. M. McEnery & G. Knowles (eds), 1–23. Harlow: Addison Wesley Longman.
- Leńko-Szymańska, A. 2014. Is this enough? A qualitative evaluation of the effectiveness of a teacher-training course on the use of corpora in language education. *ReCALL* 26(2): 260–278.
- Littlemore, J. & Oakey, D. 2004. Communication with a purpose: Exploiting the Internet to promote language learning. In *ICT and Language Learning: Integrating Pedagogy and Practice*, A. Chambers, J. Conacher & J. Littlemore (eds), 95–119. Birmingham: University of Birmingham Press.
- Lüdeling, A., Evert, S. & Baroni, M. 2007. Using web data for linguistic purposes. In *Corpus Linguistics and the Web*, M. Hundt, N. Nesselhauf & C. Biewer (eds), 7–24. Amsterdam: Rodopi.
- McCarthy, M. 2008. Accessing and interpreting corpus information in the teacher education context. *Language Teaching* 41(4): 563–574.
- McEnery, T., Xiao, R. & Tono, Y. 2006. *Corpus-Based Language Studies: An Advanced*

- Resource Book*. London: Routledge.
- Milton, J. 2006. Resource-rich web-based feedback: Helping learners become independent writers. In *Feedback in Second Language Writing: Contexts and Issues*, K. Hyland & F. Hyland (eds), 123–137. Cambridge: Cambridge University Press.
- Mondorf, B. 2007. Recalcitrant problems of comparative alternation and new insights emerging from Internet data. In *Corpus Linguistics and the Web*, M. Hundt, N. Nesselhauf & C. Biewer (eds), 211–232. Amsterdam: Rodopi.
- Nesi, H. 2000. *The Use and Abuse of EFL Dictionaries*. Tübingen: Max Niemeyer.
- O’Sullivan, Í. & Chambers, A. 2006. Learners’ writing skills in French: Corpus consultation and learner evaluation. *Journal of Second Language Writing* 15(1): 49–68.
- Park, K. 2012. Learner-corpus interaction: A locus of microgenesis in corpus-assisted L2 writing. *Applied Linguistics* 33(4): 361–385.
- Park, K. & Kinginger, C. 2010. Writing/thinking in real time: Digital video and corpus query analysis. *Language Learning & Technology* 14(3): 31–50.
- Pérez-Paredes, P., Sánchez Tornel, M., Alcaraz Calero, J. & Aguada Jiménez, P. 2011. Tracking learners’ actual uses of corpora: Guided vs non-guided corpus consultation. *Computer Assisted Language Learning* 24(3): 233–253.
- Philip, G. 2011. ‘...and I dropped my jaw with fear’: The role of corpora in teaching phraseology. In *Corpora, Language, Teaching, and Resources: From Theory to Practice*, N. Kübler (ed.), 49–68. Bern: Peter Lang.
- Quaglio, P. 2009. *Television Dialogue: The Sitcom Friends vs. Natural Conversation* [Studies in Corpus Linguistics 36]. Amsterdam: John Benjamins.
- Renouf, A., Kehoe, A. & Banerjee, J. 2007. WebCorp: An integrated system for web text search. In *Corpus Linguistics and the Web*, M. Hundt, N. Nesselhauf & C. Biewer (eds), 47–67. Amsterdam: Rodopi.
- Robb, T. 2003. Google as a quick ‘n’ dirty corpus tool. *TESL-EJ* 7(2): n.p. <<http://www.tesl-ej.org/wordpress/issues/volume7/ej26/ej26int>> (1 July, 2007).
- Rodgers, O., Chambers, A. & LeBaron, F. 2011. Corpora in the LSP classroom: A learner-centred corpus of French for biotechnologists. *International Journal of Corpus Linguistics* 16(3): 392–358.
- Rohdenburg, G. 2007. Determinants of grammatical variation in English and the formation/confirmation of linguistic hypotheses by means of Internet data. In *Corpus Linguistics and the Web*, M. Hundt, N. Nesselhauf & C. Biewer (eds), 191–209. Amsterdam: Rodopi.
- Römer, U. 2010. Using general and specialised corpora in English language teaching: Past, present and future. In *Corpus-Based Approaches to English Language Teaching*, M.-C. Campoy, B. Bellés-Fortuńo & M.-L. Gea-Valor (eds), 18–35. London: Continuum.
- Rosenbach, A. 2007. Exploring constructions on the web: A case study. In *Corpus Linguistics and the Web*, M. Hundt, N. Nesselhauf & C. Biewer (eds), 67–190. Amsterdam: Rodopi.
- Rundell, M. 2000. The biggest corpus of all. *Humanising Language Teaching* 2(3): n.p. <<http://www.hltmag.co.uk/may00/idea.htm>> (7 June, 2012).
- Scheffler, P. 2007. When intuition fails us: The world wide web as a corpus. *Glottodidactica* 33: 137–145.
- Sha, G. 2010. Using Google as a super corpus to drive written language learning: A comparison with the British National Corpus. *Computer Assisted Language Learning* 23(5): 377–393.

- Sharoff, S. 2006. Creating general-purpose corpora using automated search engine queries. In *WaCKy! Working Papers on the Web as Corpus*, M. Baroni & S. Bernardini (eds), 63–98. Bologna: Gedit.
- Shei, C. 2008a. Web as corpus, Google, and TESOL: A new trilogy. *Taiwan Journal of TESOL* 5(2): 1–28.
- Shei, C. 2008b. Discovering the hidden treasure on the Internet: Using Google to uncover the veil of phraseology. *Computer Assisted Language Learning* 21(1): 67–85.
- Sinclair, J. 2001. Preface. In *Small Corpus Studies and ELT: Theory and Practice* [Studies in Corpus Linguistics 5], M. Ghadessy, A. Henry & R. Roseberry (eds), vii–xv. Amsterdam: John Benjamins.
- Sinclair, J. 2003. *Reading Concordances: An Introduction*. Harlow: Longman.
- Sinclair, J. (ed.). 2004. *How to Use Corpora in Language Teaching* [Studies in Corpus Linguistics 12]. Amsterdam: John Benjamins.
- Sinclair, J. 2005. Corpus and text: Basic principles. / Appendix: How to build a corpus. In *Developing Linguistic Corpora: A Guide to Good Practice*, M. Wynne (ed.), 5–24 / 95–101. Oxford: Oxbow Books.
- Smith, S. 2011. Learner construction of corpora for general English in Taiwan. *Computer Assisted Language Learning* 24(4): 291–316.
- Sockett, G. & Toffoli, D. 2012. Beyond learner autonomy: A dynamic systems view of the informal learning of English in virtual online communities. *ReCALL* 24(2): 138–151.
- Stewart, D., Bernardini, S. & Aston, G. 2004. Ten years of TaLC. In *Corpora and Language Learners* [Studies in Corpus Linguistics 17], G. Aston, S. Bernardini & D. Stewart (eds), 1–18. Amsterdam: John Benjamins.
- Sun, Y.-C. 2007. Learner perceptions of a concordancing tool for academic writing. *Computer Assisted Language Learning* 20(4): 323–343.
- Todd, R. 2001. Induction from self-selected concordances and self-correction. *System* 29(1): 91–102.
- Tyne, H. 2012. Corpus work with ordinary teachers: Data-driven learning activities. In *Input, Process and Product: Developments in Teaching and Language Corpora*, J. Thomas & A. Boulton (eds), 136–151. Brno: Masaryk University Press.
- Volk, M. 2002. Using the web as corpus for linguistic research. In *Tähendusepüüdja: Catcher of the Meaning – A Festschrift for Professor Halduur Oim*, R. Pajusalu & T. Hennoste (eds), n.p. Tartu: University of Tartu. <http://www.ifi.unizh.ch/CL/volk/papers/Oim_Festschrift_2002.pdf> (25 March, 2006).
- Widdowson, H. G. 2000. On the limitations of linguistics applied. *Applied Linguistics* 21(1): 3–25.
- Willis, J. 1998. Concordances in the classroom without a computer. In *Materials Development in Language Teaching*, B. Tomlinson (ed.), 44–66. Cambridge: Cambridge University Press.
- Wu, S., Franken, M. & Witten, I. 2009. Refining the use of the web (and web search) as a language teaching and learning resource. *Computer Assisted Language Learning* 22(3): 249–268.
- Yoon, H. & Jo, J. 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.

Alex Boulton. (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, p. 267-295.
DOI: 10.1075/scl.69.13bou

Young, B. 2011. The grammar voyeur: Using Google to teach English grammar to advanced undergraduates. *American Speech* 86(2): 247–258.