ASp

la revue du GERAS

69 | 2016

Concepts and Frameworks in English for Specific Purposes

Integrating corpus tools and techniques in ESP courses

L'intégration de techniques et d'outils de corpus dans les cours d'anglais de spécialité

Alex Boulton

Electronic version

URL: <http://journals.openedition.org/asp/4826>

DOI: 10.4000/asp.4826

ISSN: 2108-6354

Publisher

Groupe d'étude et de recherche en anglais de spécialité

Printed version

Date of publication: 9 March 2016

Number of pages: 113-137

ISSN: 1246-8185

Electronic reference

Alex Boulton, « Integrating corpus tools and techniques in ESP courses », *ASp* [Online], 69 | 2016, Online since 01 March 2017, connection on 02 November 2020. URL : <http://journals.openedition.org/asp/4826> ; DOI : <https://doi.org/10.4000/asp.4826>

This text was automatically generated on 2 November 2020.

Tous droits réservés

Integrating corpus tools and techniques in ESP courses

L'intégration de techniques et d'outils de corpus dans les cours d'anglais de spécialité

Alex Boulton

Introduction

- 1 At its most basic level, corpus linguistics offers us a range of tools and methodologies to find out about language. In many areas, they have become so ubiquitous that it is hard to imagine work without them – including in English for Specific or Academic Purposes (ESP/EAP). Given their multiple affordances (Leńko-Szymańska & Boulton 2015), users in different disciplines understandably tend to interpret and exploit corpus linguistics tools and methodologies in different ways, depending on their specific needs, aims and contexts. Even defining a corpus is not uncontroversial among corpus linguists (Gilquin & Gries 2009), and determining the scope of corpus linguistics can generate heated debate (Worlock Pope 2010). A prototypical description such as ‘a large collection of authentic texts representative of a target variety in electronic format’ is a rather catch-all definition designed to appease many users. This in itself highlights the interest of a corpus approach, which allows us to get computers to help convert language *data* from relevant texts into usable and useful *information* about frequent meaning and use in context (Hanks 2013). In other words, corpus linguistics is not concerned with what is *possible* in a language, but in what is *probable*.
- 2 Experts in corpus linguistics compile and analyse vast quantities of texts to further our understanding of language and fine-tune descriptions, and have informed a wide variety of reference materials relevant to teaching and learning foreign or second languages (L2), the most visible being dictionaries, grammar books and usage manuals. Corpora can also underpin syllabuses, testing and assessment, as well as coursebooks and other materials (cf. McCarthy 2004). They have been widely used in EAP and ESP to derive frequency lists of words and phrases (e.g. the Academic Word List – Coxhead

2000; the New Academic Vocabulary List – Gardner & Davies 2014; the Academic Formulas List – Simpson-Vlach & Ellis 2010; or the Academic Collocations List – Ackermann & Chen 2013), and to improve descriptions of particular genre- or discipline-specific terminology and discourse (e.g. Boulton et al. 2012).

- 3 There are two main implications of this from the point of view of language teaching and learning. First, though improved descriptions of language can only be applauded, the end-users may be entirely unaware of the corpus input and end up being “consumers” where they could be “active participants” (McCarthy 2008: 565-566). Using corpora to help with language description may have no impact on teaching practice in terms of the activities covered or the respective roles of teachers and learners. Second, general-purpose tools can only provide general-purpose answers. This is not to criticise such resources, which often provide the relevant information quickly and easily, but merely to acknowledge their inevitable limitations (Frankenberg-Garcia 2014). The goal of any ready-made resource is to simplify and generalise in order to satisfy a maximum number of users in as short a space as possible, a job they do remarkably well. Nonetheless, even resources that describe seemingly specialised varieties such as medicine or research articles cannot reasonably be expected to address every conceivable language point: a biochemist in immunology may find that even corpus-based resources do not contain answers to his or her specific questions – what does this word or phrase mean in my field, how is it used, is it worth remembering?
- 4 We begin with a brief overview of “data-driven learning” (DDL) in its wider sense before homing in on ESP. The bulk of the paper is designed to demonstrate the basic concepts through specific examples of the sort of things ESP learners can do with different types of software and data, accompanied by step-by-step instructions. Though these are discussed, explained and contextualised in some detail, the hope is that they will be more accessible than abstract, theoretical descriptions. All of the examples given here derive from genuine queries conducted by, with or for actual learners in response to their own individual queries or problems. Crucially, they all use free and widely-available resources from the internet. This sets them apart from much DDL work, which may be seen as a theoretical undertaking appropriate for high-level, linguistically sophisticated learners enrolled in language degrees, a view cast into doubt twenty-five years ago (Johns 1991: 12).
- 5 First, in a weak but potentially highly accessible form, web searches involve querying vast quantities of language; while many learners may already be doing this informally, a little training and consciousness-raising may bring substantial and immediate benefits. The underlying philosophy is not dissimilar to DDL, so this may provide a lead-in to some of the more user-friendly online tools such as those provided by Davies on his website at Brigham Young University (BYU).¹ We then move on to generic corpora such as the British National Corpus or the Corpus of Contemporary American English, which are sufficiently large and well-structured to allow quite narrow searches for specific purposes. Nonetheless, users with specific long-term needs may wish to go further and create their own dedicated corpus; in our third set of examples, familiarity with the contents of such small, personal corpora directly in relation to individual needs can help with formulating queries and interpreting the results. We end the paper with a brief overview of findings from DDL research to date.

1. Data-driven learning

- 6 Though not every user can have a tailor-made set of reference resources, one lead may be to look at the processes involved in creating such resources and see if learners can act in similar ways to create their own. In other words, rather than relying on experts to pre-digest the language, maybe learners could chew on the language data themselves. This is what Johns (1990) christened “data-driven learning” which he characterised as “the attempt to cut out the middleman as far as possible and to give the learner direct access to the data, the underlying assumption being that effective language learning is a form of linguistic research” (p. 18). The term itself is perhaps intentionally provocative (Boulton 2011a), and Johns’ characterisation of DDL as “radical” (1988: 20) and “revolutionary” (1990: 14) may have been intended to attract attention and prompt rethinking of contemporary knowledge-transmission teaching practices. This might have seemed necessary at the time, but the downside is that DDL is often perceived as a threatening, even scary concept. An alternative might be to build bridges with familiar practices: if, in essence, DDL is perceived as little more than encouraging learners to take a bit of initiative to explore language and figure things out for themselves, it might find a more receptive audience among teachers already attempting essentially the same thing in their own practice (Boulton & Tyne 2014). It might appeal to those who are keen to return language to a central place in their language class, and rather than expecting teachers to make the conceptual leap towards corpus linguistics, it may help to bring DDL closer to them by highlighting how DDL exploits any number of key concepts in existing approaches – including, but not limited to: authenticity, autonomy, cognitive depth, consciousness-raising, constructivism, context, critical thinking, discovery learning, heuristics, ICT, individualisation, induction, learner-centeredness, learning-to-learn, lifelong learning, (meta-)cognition, motivation, noticing, sensitisation and transferability.
- 7 Certainly DDL would seem to be in line with much of what we know about language and processing. Usage-based theories (e.g. Tomasello 2005) suggest that we need massive exposure to language, but naturalistic contact is simply too rare, especially in foreign-language contexts (e.g. Schmitt et al. 2016). Zahar et al. (2001: 558) calculate that with an hour of reading a week, their students would need 29 years to acquire 2,000 words incidentally from that reading; DDL can help to organise and focus the exposure (Gaskell & Cobb 2004). Language is not rule-driven but fuzzy and probabilistic in nature (Hanks 2013), with grammar and meaning both emerging from use (Beckner et al. 2009). And the mind works with exemplars beyond the level of word in line with dynamic systems theory (Larsen-Freeman & Cameron 2008), Sinclair’s (1991) idiom principle, Hoey’s (2005) lexical priming or Taylor’s (2012) model of the mental corpus, and finds support in recent psycholinguistic work on ‘chunking’ (e.g. Millar 2011), among others.
- 8 Human cognition itself is based on pattern detection, which evolutionary psychology explains is a remarkably adaptive system (e.g. Barrett et al. 2002). However, it is not perfect: we often see patterns and find connections that are not really there. For example an English learner of French, on being offered a *noix*, might assume this refers to the entire class of ‘nuts’ rather than just walnuts. And once beliefs are sufficiently well established, they become remarkably difficult to dislodge: a learner of English who regularly uses *discuss about* for whatever reason can be remarkably impervious to

copious input suggesting that there is usually no preposition involved. Noticing is thus crucial in language learning (Schmidt 1990), but often needs a helping hand, which is where DDL comes in.

2. Parallels to DDL in everyday life

- 9 McCarthy (2008: 566) claims that “we are, all of us, corpus users, because we use the internet.” While this is controversial in some quarters, one does not need to accept the status of the web as a corpus – still less accept a search engine as a concordancer – to see parallels between corpus queries and web searches. Both involve using software to help turn language data into useful, relevant information, and many learners are already using search engines for purposes explicitly related to their language questions (Geluso 2013). Informal use of Google for language learning purposes seems to be extremely widespread (Conroy 2010), but has as yet received very little attention in terms of research, even in language for general purposes. If one road to integrating DDL in the ESP class is to build bridges with ordinary practice (the specificity deriving from the search queries rather than the corpus itself), then this may bear further exploration (Boulton 2015).
- 10 To the extent that web searches resemble corpus queries, this may be exploited in two directions. Most obviously, we might seek to build on existing skills and techniques to bring learners towards DDL; less intuitively, we might see if we can bring DDL closer to the learner by emphasising search techniques that most closely resemble Googling. In either case, learners do tend to see the parallels of their own accord (Sun 2007), and initially at least tend to approach corpus consultation by applying similar techniques to those they use for web searches (Pérez-Paredes et al. 2012). The transfer of procedures suggests a porous boundary between the two which can be exploited in other ways: the training required for the transition to corpus consultation might be implemented gradually by helping learners in initial stages to improve their web search techniques. From a purely DDL perspective, the further the learners progress before they even encounter a concordancer, the smaller those final delicate steps will need to be.
- 11 More pragmatically, though we would of course hope that some learners will continue corpus consultation after the end of a course, this is unlikely to be the case for all. All things being equal, the more specialised the tool, the less frequently it will be used, which will in turn lead to a loss of efficiency and thus trigger a vicious spiral leading to ultimate non-use. For many learners, the long-term benefits of training will thus be lost. Conversely, since learners are already using Google frequently (including for language learning) and will continue to do so between classes and after the end of their course, any improvement in their search techniques and interpretation skills is more likely to persist in the long term.
- 12 Informal surveys among my own students suggest most of them have very limited ideas of how Google works or of the range of options it offers. The joy of such tools is that immediate benefit can be derived without any training at all, just as no training is required to use a dictionary. But as with a dictionary, a little training is likely to dramatically improve the results (Nesi 2000). At an abstract level, learners can find it an eye-opener to learn how Google collects and ranks web sites, or interprets queries in the light of search histories to personalise the output. At a more practical level, a number of advanced search functions can be extremely useful in L2. Filters include

language (to discriminate cognate search terms) and region (to isolate American or British English), last update (to exclude older data), site or domain (to limit to a free online academic website or .ac.uk, for example), and format (limiting results to .pdf files may be more appropriate for some queries). Boolean searches can be useful to exclude some words (e.g. – *bird* in an engineer’s search featuring *crane*) or to force Google to use the exact search query rather than over-interpret it, a function which can be achieved simply by encasing the entire query within inverted commas – “one of the best tips to teach your students” (Dudeney 2000: 22). This can be combined with the only wildcard that Google allows, the asterisk to signify any word within the phrase as in the following example:

“play a * role in”

- 13 This example (see Figure 1) is motivated by a learner’s repeated use of *important* within a single assignment – a common “lexical teddy bear” (Hasselgren 1994) which learners feel comfortable using frequently in a wide range of contexts. The query returns the usual presentation of short extracts or snippets (akin to the concordance lines produced by corpus software), but the word in the place of the asterisk can vary. Figure 1 shows the unedited results for the first few hits; though these should of course be interpreted with caution, they provide a valuable source of information to enrich productions which may be difficult or impossible to obtain from traditional resources. A thesaurus will of course provide a list of synonyms, but without context will be far more difficult to interpret than this minimally contextualised Google search – here *play a [key, critical, significant, important, leading, major] role*. The critical thinking involved here in turning the real question into a query that Google can understand, and sifting the results to arrive at a relevant, usable conclusion, are no less crucial in corpus consultation proper: sensitising learners to such issues at the level of web searches is in itself and may help make DDL more accessible later on.
- 14 Internet search engines are by no means the only everyday parallels to corpus tools or use of corpus-like techniques. A simple CTRL+F search command within a document or web page highlights repeated occurrences of a target item that can be scrolled through easily; the online bookstore Amazon offers a “search inside” feature for many books which not only highlights the target word(s) but also provides snippets in a sidebar (as does Google Books, or some versions of MS Word). Many learners are also familiar with Linguee, which offers human translations of a variety of documents in many European and other major languages. It poses a certain number of problems since it changes from day to day, and we have no way of sorting the items or even knowing which was the source and which the target; nevertheless, it can have its uses. For example, the Oxford Advanced Learner’s Dictionary of English gives one meaning of *reflect* as “to think carefully and deeply about something”, i.e. a possible translation of *réfléchir*. Typing *réfléchir* into Linguee, the first twenty returns in English do indeed include five occurrences of *reflect*, but thirteen of *think/thought* (plus 2 other translations). The other way round, *reflect* gives fifteen occurrences of *réfléter/reflet* and only two of *réfléchir* (plus 3 other translations). In other words, though a possible translation of *réfléchir*, *reflect* seems to be relatively unusual in this sense, and the learner may be better advised to choose an alternative – information impossible to get from the dictionary entry (cf. Boulton & De Cock forthcoming). Again, despite its limitations and the need for careful thinking in interpreting the results obtained, it is an example of how everyday, familiar technology can be used to extract information from language data; a little training can be a useful end in itself, and may make DDL more accessible at

a subsequent stage. Other tools such as WebCorp allow more linguistically sophisticated queries and output from the web, and may also serve as a springboard to hands-on corpus work.

Figure 1. Google hits for “*play a * role in*”

Floral volatiles **play a key role in** specialized ant pollination
www.sciencedirect.com/science/.../S1433831913000747 ScienceDirect
 by C de Vega - 2014 - **Related articles**
 They can affect the structure and functioning of terrestrial ecosystems and **play a significant role in** ecologically different habitats from tropical forests to ...

A Gene Linked to Disease Found to **Play a Critical Role in** ...
www.scripps.edu/.../e.../puthanveetil.html Scripps Research Institute
 Aug 25, 2014 - A Gene Linked to Disease Found to **Play a Critical Role in** Normal Memory Development. By Eric Sauter. It has been more than 20 years since ...

Everyone Can Play a Role in the Conversation about Mental ...
store.samhsa.gov/... Substance Abuse and Mental Health Services Admini...
 Faith and community leaders can **play a significant role in** helping to educate individuals and families about mental health. This fact sheet can help communities ...

Can the elderly play an important role in today's society? | A ...
www.ted.com/conversations/.../can_the_elderly_play_an_import.html TED
 Do those elderly **play a large role in** today's society that is not apparent to us? If not, what largely influential role could we give them that could ultimately change ...

Gene linked to disease found to **play a critical role in** normal ...
www.sciencedaily.com/releases/2014/.../140812235302.html Science Daily
 Aug 12, 2014 - It has been more than 20 years since scientists discovered that mutations in the gene huntingtin cause the devastating progressive ...

A strong voice for global sustainable development: how ...
www.die-gdi.de/.../a-strong-voice-for-global-sustainable-development-h...
 A strong voice for global sustainable development: how China can **play a leading role in** the Post-2015 Agenda. A strong voice for global sustainable ...

Immigration Could Play a Role in Arkansas Senate Race ...
www.usnews.com/.../immigration-could-play... U.S. News & World Report
 Aug 4, 2014 - Therefore, I think immigration is not going to **play a major role in** most Senate contests." Cotton Hits Sen. Pryor On Amnesty To Win Swing ...

Fisheries and aquaculture **play a key role in** Africa's economy
www.fao.org/fishery/news/40619/en Food and Agriculture Organization
 Jul 11, 2014 - The fisheries and aquaculture sector contribute significantly to Africa's overall economy. In a new study published by FAO, the value added by ...

3. General-purpose corpora

- 15 Should the teacher decide that the students' needs justify the investment, a relatively straightforward next step might involve the type of large corpus designed to be representative of a given language as a whole and which therefore covers a wide range of needs. Usually quite carefully compiled and balanced, many can be queried on line from a single integrated webpage (no download necessary) using a simple interface designed partly with non-corpus linguists in mind. For English, one of the most widely used such sources is the BYU interface to a range of varieties of corpora: American, Canadian and British English, as well as GloWBE for twenty varieties from inner- or outer-circle countries, and corpora from soap opera transcripts to historical texts and Wikipedia. The platform is entirely free (though registration is required), stable, reliable, unencumbered by advertisements or spam, and relatively straightforward to use. Among the most frequently consulted by users who declare their primary role as language teachers or learners are the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA), which will serve as examples here for British and American English respectively. The BNC is a static corpus, while COCA is updated regularly (searches can be limited to recent years if necessary), but they are otherwise very similar in most major respects and use essentially the same interface. COCA's architecture is modelled on the BNC, being divided into similar sub-corpora (e.g. SPOKEN

or ACADEMIC), and searches can be further refined to another level of granularity (e.g. SCIENCE AND TECHNOLOGY or the HUMANITIES). However, Davies does not have the resources available to the BNC consortium and relies on semi-automated procedures for harvesting data from the internet; this is partly compensated by COCA's larger size (450 million words compared to the BNC's 100 million), though care will be needed in interpreting any corpus data (Maniez 2012).

- 16 Rather than providing abstract descriptions, a couple of queries resulting from genuine learners' needs may provide a glimpse of some of the affordances of such corpora in ESP/EAP. If sports students have difficulty using an appropriate verb (*play* or *do*) with different activities, authentic examples can be manually searched for, selected, appropriately formatted and printed out. Figure 2 shows the typical KWIC format (key word in context) that may seem disconcerting on first encounter, but it helps to focus attention on the target item which is centred in bold, and realising they do not represent entire sentences reduces the tendency to read for content and an attempt at understanding every word. The task here is to identify the verbs on the left, sort them, and try to find patterns in the meaning – i.e. what distinguishes the two groups of words. A teacher who knows that sporting activities ending in *-ing* collocate with the verb *go* may teach this seemingly simple rule only to find that students have difficulty applying it, perhaps in part because it is so simple. Requiring learners to actively sort the examples, “*linguaging*” their findings in pairs (Swain 2006), can lead to greater depth of cognitive processing, thus enhancing retention. DDL work is often claimed to place considerable demands upon the teacher, and to be mostly suitable for advanced, motivated and linguistically sophisticated learners. However, an activity such as this is not particularly time-consuming to create, and can be used successfully with learners at lower levels of proficiency by virtue of judicious selection of the lines for focused language content, and provision of clear instructions and guided tasks in a simple hand-out (e.g. Boulton 2009).

Figure 2. Selected concordance lines for [*play*]/[*go*] + [sporting activity] from the BNC

rday afternoon?" "The senior girls play **badminton**; others play hockey or netball." "Well, d
all week. Oh I It's alright. Do you play **basketball**? Is there a family history of being very
wn. Who's her? So Mrs. She was going **climbing** Oh oh! in the Lake District. and looked I
d stopped moaning and learned to play **cricket** like we Pakistanis --; properly! Then they m
we loved it there. Well you used to go **fishing** there didn't you? Yes In (unclear) er one of
at Robinson College, said: "I've played **football** since I was 10 or 11. It's been a bit of a
en 1m and 2m people who want to play **golf** but can not get into a club," says Neil Hayward
ub's east wall was also used for playing **handball**, he says "like squash but without the racke
rts ground. There you will learn to play **hockey** or rugby. Later you may be able to choose
also plays the piano and harp and goes **horse-riding**. IT WAS a first-time success for SAG
fty arrowheads. When the Indians went **hunting**, whether for animals or for rival Indians, t
flamboyancy. "We simply did not play **rugby** at Strasbourg," explained Dubroca. "Our play
o reach it. We took our time. We went **sailing** and swimming, and once we went out at nigh
their children to go on them. If they go **skiing** they might break their leg ... Parents have to
olarships to study for degrees and play **soccer** in Pennsylvania. Ian, 20, of Briar Walk, is t
urgh. Claire, who is 17, started playing **squash** when she was ten years old at Ellon Squash
rafts, but she hates to get changed to go **swimming** or playing games in the gym or playing g

- 17 There are two obvious objections to this. First, there is no guarantee that learners will come to the correct conclusion. The rejoinder is that if learners do not remember the correct rule as traditionally given, then any even partially correct answer is better than nothing and can subsequently be refined – Aston (2001) sees such learning as a series of gradually refined approximations to the target. Further, a rule which learners come up

with themselves is more likely to be meaningful to them and relevant to their own interests (linguistic or otherwise), a view compatible with constructivism (cf. Cobb 1999 for a corpus-inspired take on this). In this particular example, learners are more likely to come up with their own suggestions for linking *go* with activities that predate modern leisure, occur outdoors, have no limited pitch or terrain, are important for survival, involve travel from one place to another, and so on; *play*, on the other hand, is typically a modern construct where the only objective is the game itself, often involving a ball, team work, competition, a pitch, etc. Despite their limitations, suggestions such as these are remarkably sophisticated and imaginative, as well as unpredictable; encouraging learners to make sense of language in this way is likely to lead to increased language sensitivity. The teacher's role, especially in initial stages, is not just to provide the language but to help refine the patterns detected, signal obviously erroneous ones, suggest alternatives including the right answer where appropriate – in the present case, simply thinking about the use of the *-ing* form can lead to a genuine and salient “a-ha moment” and promote further learning (Larsen-Freeman & Cameron 2008: 59).

- 18 A second objection is that the process takes considerable time for relatively little result (Thompson 2006). This is undoubtedly true to an extent; for the most immediate short-term effects, simply giving the students the essential information in a traditional knowledge-transmission model can be highly effective (Kirschner et al. 2006). The advantage of DDL is that the processes may take some time to begin with, but with practice become much quicker and more sophisticated, and help the learner to become more linguistically aware and autonomous in dealing with authentic language (Boulton 2010b), though of course such things are extremely difficult to test empirically (Boulton 2012b) given the number of variables involved (Larsen-Freeman & Cameron 2008).
- 19 The use of printed concordances or other corpus data may lead to hands-on corpus work (Boulton 2012c) such as in a second example here. The simplest queries with the BYU interface, as with most other corpora, consist of simply entering one or more words and pressing ENTER. For example, let us return to *important*, which is not just a lexical teddy-bear but is also used differently in English and French. A master's student writes of *an important number of studies* in her field, a formulation which the teacher feels uncomfortable with, and guides the student in using COCA to find out why. If the student has not registered to use it, or has no previous experience of the BYU corpora, the teacher can perform the searches automatically generating a separate URL for each query (here <http://corpus.byu.edu/coca/?c=coca&q=41635465>). This can then be emailed to the student who is guaranteed to see the same thing as the teacher – both query forms and results. Alternatively, the instructions might run as follows:
 1. Open COCA and in the WORD(S) box, type *important number*. Click SEARCH (or press ENTER).
How many occurrences of *important number* are there in COCA?
Is this a lot or not?
 2. Click on the result for *important number* and look at the concordance lines.
What numbers are qualified as *important*?
Is the meaning similar to what you want to say in your context?
 3. Reset, select the ACADEMIC section to limit your search. In POS LIST (part of speech), select ADJECTIVE followed by *number* so that your query looks like this: *[j*] number*.
What adjectives typically precede *number*?
Are there any that might be appropriate for what you want to say in your context?

- 20 The set of queries (or “transactions”, to use Park and Kinginger’s 2010 terminology) becomes increasingly sophisticated, highlighting some of the different functions and reasoning involved. In Step 1, frequency is a slippery concept and may need discussion. The phrase *important number* occurs thirty-two times in the 450M words of COCA; if the average book sold by Amazon contains 64K words, then you would expect to encounter *important number* just once in about 220 books. This would probably seem lower than most French speakers’ intuition for *nombre important* in their own language; so though it may not be impossible in English, it certainly seems to be relatively unusual. In Step 2 (Figure 3), the specific numbers qualified as important are 9.5, five, 350, one, twelve, 7.5%, four, 13%, and eight; tellingly, 13% is described as “a small but important number”. In other words, *important* doesn’t simply mean big in English as the student originally intended, a conclusion which she might have difficulty in deriving from dictionaries (which include such definitions as “of great value”) or in noticing through chance encounters with the word in context, especially in the absence of negative evidence. In Step 3 (Figure 4), *important* does not make it to the list of the twenty most frequent adjectives preceding *number*; its frequency of just nine puts it well behind the more usual choices with the desired meaning here such as *large*, *substantial*, *high*, *great*, and so on, which the student might decide to settle on in her writing.

Figure 3. The 32 concordance lines for *important number* in COCA

the Labor and Liberal parties, and including an **important number** of senior Tories, it took the name Arms and ti
think the number-one reason, the most **important number**, and in some ways the only one that matters is 9.5.
against the backdrop of what I consider the most **important number** in all the blizzard of numbers we're hearing
ll said. " Right now, the **important number** to focus on is, since the 2006 agreement was struck, we have
ral had kept all of them captive. Five was an **important number**, one that Ling Yun had taken for granted until F
s bemused rather than impressed. " It's an **important number** for us computer jockeys. " Paul thought a little lo
ntage of the Latino vote " the single most **important number** " to come out of the election. Earlier this year, the
affinity groups, and 30,000 events. But the most **important number**, and the clue to how Obama's machine bas
took 350 to be the most **important number** on the planet, what would it imply? In essence, it means
ve have a number-350. The most **important number** on earth. If the Internet has a cosmic purpose, this could b
entists call the transfer efficiency. It's an **important number**, as it gives us some idea about how many meteorit
nstitution passed. And of course, that other **important number**, how many Iraqis turned out to vote -- Tony? HA
: leaving the labor force, but the really **important number** is how many people are out there getting new jobs. TI
823 people... Mayor BLOOMBERG: The most **important number**, though, is that we did not lose one life during th
everything else in the budget. The most **important number** in the CBO report says that if you measure the non-d
nager, the Almost Head Honcho. Twelve was an **important number**, Mack decided, with its own name. A dozen.
ld yield the pulsar mass, an extremely **important number**. " But he figured the odds were against him. All the pu
makes it easy to obtain the most **important number** for buying a new car: the dealer cost, or invoice price.
so vast because there is one crucially **important number** in nature that is equal to **29;1411;TOOLONG, 000,00
ive and well in education, that's an **important number** for us, " says Dixon, referring to the 7.5% figure. "
ig about Roy's record is it deals with the most **important number** on the stat sheet: winning. Not even included i
r less food-deprived overall. " The most **important number** on a nutrition label is serving size, says Allred. " A k
. pp. 197-200). Four is an **important number** that appears in stories and traditions throughout the world. In one
d an optical interferometer of the FabryPerot type. An **important number** of these instruments have been design
n Hollywood, Fla. " This is a very **important number** to watch because once it starts falling from this level you've
Carlos Reservation, Gila County also provides an **important number** of Apache defendants. 12 Pima and Pinal re
rst five propositions were submitted to Rome by an **important number** of French bishops. All the doctors could d
l commercial radio stations that have included an **important number** of programas folcloricos in their broadcasti
ch power " were voiced by 13%, a small but **important number** of responders. Conclusion ≠ To our knowledge,
. I think that it is an **important number**, surely, but I think we've got a moving target here,
ntrary to Ferretti's erroneous assertion that eight is the most **important number** (1975:24), the Kom consider od
the four periodicals, this overall aggregate score is an **important number**. Thus, these four representative publi

Figure 4. Query details and the 20 most frequent adjectives preceding *number* in academic texts in COCA

4. Purpose-built corpora

- 21 The BYU corpora have all the advantages of general-purpose tools: they are large, professionally compiled, and suitable for many purposes. However, this is also potentially their failing for some specific purposes. A number of specialised corpora are available for use on line (e.g. the Business Letters Corpus; the British Law Report Corpus via LexTutor; or the Michigan Corpora of Academic Spoken English/Upper-level Student Papers); others can be downloaded (e.g. the British Academic Written/Spoken English corpora, both available from the University of Oxford Text Archive). Nesi (2015) provides a useful rationale for small, manually-sampled corpora, rich in contextual information, for research, teaching and learning in English for Specific Purposes (ESP) and English for Academic Purposes (EAP). Even here though, the user may find that the texts selected for inclusion do not reflect their specific needs closely enough. In such cases, the teacher or learner may decide to compile their own corpus.
- 22 One of the quickest ways to compile a corpus is with a tool such as BooTcAT. Inputting a handful of seed items – i.e. words chosen to be specific to the target field – will bring up webpages that contain combinations of these in just a few seconds. Various options are available for types of pages to include or exclude, and individual pages can be deleted or added as required. More usually, however, local corpora are built by manually selecting texts for use with free downloadable software such as AntConc. This can be used for serious research in corpus linguistics, and a number of additional tools are available on the website. But AntConc itself has a streamlined interface in English (Figure 5) and is accompanied by a series of short video tutorials for the main tools which makes it suitable for use with very little training among non-linguistics students, as evidenced in a number of studies to date (e.g. Charles 2012). Having students build their own small corpora can also lead to a sense of ownership and familiarity with the contents (Smith 2011).

Figure 5. The AntConc interface

- 23 In the corpus compilation stage, a number of decisions need to be made about the types of texts to include from different sources, the number of texts and the overall word count, and any editing to be carried out. There are no blanket answers to such questions. A simple corpus might consist only of a single textbook, a business report, or a dozen research articles in .txt format. While more and cleaner texts may be theoretically desirable, the specific decision will depend on the individual's needs and objectives, as well as their technical know-how and the time they are prepared to invest. If they can gain the information they need from a small, relatively dirty corpus, then that will be enough (Aston 1997). Again, rather than discussing the procedures in abstract terms, the rest of this section will outline the type of query learners can usefully perform to answer their own questions using AntConc. For present purposes, the corpus here consists of 110 published academic papers that evaluate some aspect of corpus use in language learning and teaching. This is highly specialised and rather larger than most learners would want to produce, and has been cleaned so that it includes mainly the authors' original text (deleting headers and footers, appendices, reference lists, tables and figures, long quotations and lists of examples, etc.), but it will serve to demonstrate the basic procedures involved for ESP users.
- 24 Once the text or corpus has been opened via the FILE menu, it is often as well to get to know what it contains. Choosing the WORD LIST tab simply counts all the words in the corpus (here 613,708 words) and ranks them in descending order of frequency. The most common items are *the*, *of*, *to*, *and*, *in*, which may not appear very meaningful (though see Gledhill 2000 on their potential importance); but even scrolling down the top twenty words we already come across some with lexical content: *students*, *corpus*, *language*. It is possible to apply a stoplist so that selected items are automatically excluded, and a lemmatiser so that different forms of the same word are counted

together, e.g. *student* and *students* (free tools are available on the AntConc website). But without any tagging or markup it is still quite easy to scroll down and identify the top ten lemmas: *student*, *corpus*, *use*, *learn*, *language*, *learner*, *word*, *study* and *write*. These items alone account for over 13% of all lexical words in the corpus, and give a fairly accurate picture of the contents – what Scott and Tribble (2006) call its “aboutness”. They are also words which the user will encounter frequently in this type of text, and will probably want to know extremely well for productive purposes too. An alternative is to create a KEYWORD LIST to compare the words in this corpus to, say, one of student writing or a larger corpus of general English; this will highlight the words which are significantly more (or less) frequent in the target corpus and which may thus be worth further investigation.

- 25 Of course, text is more than a list of words, and it is essential to see how words group together as chunks, clusters or bundles. Similar processes are involved in the creation of such lists by choosing CLUSTERS/N-GRAMS, and selecting the number of words in each string; it can also be useful to set a minimum range, i.e. how many different texts the string must appear in. For example, the most frequent bi-grams (two-word sequences) are: *of the*, *in the*, *to the*, *the students*, *on the*. As before, these may not carry much meaning, but can lead to discussion of the importance of prepositions and determiners. Looking at longer stretches such as 10-grams shows that very few occur in more than one paper, and are almost invariably quotes that are attributed to an original source, e.g. Johns (1991: 2) positing the language learner as a “research worker whose learning needs to be driven by access to linguistic data.” This can lead to discussion of where intelligent plundering of the language for genre-appropriate recyclable phrases (patch-writing) becomes genuine plagiarism.
- 26 The strings likely to be most useful to students tend to be of intermediate length; Byrd and Coxhead (2010) identified four-word bundles as particularly helpful across disciplines in academic writing. The most frequent in our corpus are given in Table 1; several of the 4-grams are not phrases in the traditional sense, insofar as they have no psychological validity, since of course the computer does not understand the language it is retrieving. Top of the list is *on the other hand* with 100 occurrences in 49 texts. Students can benefit from noting that nearly half of all texts in our corpus use it, twice each on average, and taking that as typical usage. It is notable too that *on the one hand* does not appear on this list, occurring only 24 times (and there are no hits at all for *on one hand* here), suggesting that the two forms do not have to go together. Even for such small words as prepositions, learners might benefit from noting specific formulations such as *at the same time* (rather than *in the same time*). The sequence *students were asked to* might also seem surprising to speakers of French, since the grammatical structure does not calque directly between the two languages and it might not occur spontaneously to them or their teachers. The fact that it is present in over a third of all the texts here – on average twice in each of them – can serve to bring their attention to an item that might otherwise be overlooked. Scrolling further down the list reveals many more potentially useful strings and points for discussion, reflection or follow-up searches.

Table 1. The most frequent four-word sequences

rank	freq	range	4-grams
------	------	-------	---------

1	100	49	on the other hand
2	84	50	the end of the
3	84	43	the use of the
4	83	32	the use of corpora
5	81	45	at the end of
6	81	38	in the case of
7	79	39	students were asked to
8	75	36	in the context of
9	73	36	the results of the
10	72	39	in the form of
11	63	31	the students in the
12	55	26	with the help of
13	53	37	the british national corpus
14	52	33	in the use of
15	49	29	at the beginning of
16	46	28	the beginning of the
17	45	32	at the same time
18	44	32	to the use of
19	43	32	the fact that the
20	42	18	in the present study

- 27 The CLUSTERS/N-GRAMS function presents only fixed sequences; greater flexibility may be sought using the COLLOCATES tool to specify any words that occur within a given span left and/or right of a chosen item. For example, within three words of the search term *learner*, the corpus contains (in addition to grammar-function words) *corpus* and *corpora* (13 occurrences to the left, 105 occurrences to the right, suggesting *learner corpus/corpora*); *language* (49L, 17R for *language learner*); *autonomy* (1L, 46R); and *English* (4L, 17R). Lower down the list but still occurring at least ten times within three words left or right of *learner*, we also find *data* (19); *learning*, *centred* (16); *teacher* (13); *individual*, *analysis*, *access* (12); *research*, *DDL* (11); *researcher*, *performance*, *NS*, *approach*, *advanced* (10). Again, all of this tells us more about how the target item *learner* is used in context

in this corpus, and the various combinations can be followed up with further searches just by clicking on them (see CONCORDANCE below).

- 28 The CONCORDANCE PLOT shows the position of a target item within each text. The literature review section which occurs early in papers is likely to take a historical perspective, as witnessed by the distribution of items such as *1980s* indicated by the vertical lines towards the left of the bars in Figure 6. A corpus of research articles divided or marked up according to IMRAD (introductions, methods, results and discussion) sections can reveal different distributions or uses of the same items. In the CONCORDANCE PLOT tool, clicking on one of the horizontal lines brings up the FILE VIEW function so that the item can be seen in the context of the text as a whole.

Figure 6. Concordance plot for *1980s*

- 29 The final tool (CONCORDANCE) provides the now familiar KWIC presentation which is at the core of DDL, hence its position as the first tab within AntConc. This is typically used for lexicogrammar as we saw in the examples from the other software, but it can also be helpful in exploring discourse and content. A cursory glance on the immediate left of *e.g.* (Figure 7) shows that it is typically used in brackets – for a total of 485 out of 571 occurrences, approximately 85% of cases; in others, *e.g.* is still within brackets but preceded by *see* (lines 2 and 14). Students can easily spot that *e.g.* is rarely used within the syntax of the main sentence. Looking on the right, it is striking that it is often used to introduce references. This highlights typical citation practice in applied linguistics in English, with the author saying what s/he means and relegating the supporting reference to the end of the clause or sentence – as opposed to beginning a sentence with *X says that...*, which novice researchers often tend to overuse in their scientific writing in English. Further searches specifically for this might include *19** or *20**, where the asterisk stands for any characters.

Figure 7. Sample concordance lines for e.g.

1 newspaper, topic-oriented) and objectives (e.g. structural, lexical, cultural) based on l
 2 learning theorists and practitioners (see, e.g., Kiraly 2000). The attention devoted to
 3 ed, thus eliminating a number of variables (e.g. Stevens 1991). The sources used are
 4 based materials have often been discussed (e.g. Boulton 2010, forthcoming), but never
 5 s dealing with highly specialized subjects (e.g. Polyester media development for inkjet
 6 reen adjacent points on the scale as equal (e.g. very much and much vs. much and
 7 sible variations in RAs across disciplines (e.g., Anthony, 1999; Hyland, 2000; Nwogu,
 8 'y) while "T conveys some but less strength (e.g., possibly, tend to). (Hyland, 1998;
 9 divided into different semantic categories (e.g. similar lexical meanings), users can
 10 guage of students, at least one university (e.g. University of South Australia, 2007),
 11 pecific and uninterrupted word sequences (e.g. a verb and an associated preposition).
 12 t; 2. music and pictures; 3. paralinguistic: e.g. choice of typeface and letter sizes; 4. f
 13 empirical base for corpus-based learning (e.g., Bowker, 1999, study of the effect of
 14 research methods in corpus linguistics (see e.g. Davies 2004; Kirk 2002; Mair 2002),
 15 , locally authenticated corpus annotations (e.g. Braun 2006). The IFAConc approach
 16 onal print materials or instructional aids (e.g., Sun, 1999). A concordancer one, type
 17 although what follows can be personalized (e.g., in this paper we present. . .) or impers
 18 doblems during concordance-based tasks (e.g. Turnbull and Burston 1998; Bernardini
 19 for second and foreign language learners (e.g., Kramsch 1997; Rampton 1990; Train
 20 cles, and missing verbs. 2. lexical errors, e.g., wrong or inappropriate word

- 30 A final example (Figure 8) derives from apprentice writers' tendency to want to prove their ideas; their teachers may find it useful to be able to provide some hard data in support of the injunction to beware of this.

Figure 8. Concordance of *prov* that*

1 ggest that they do have something to offer, 2 tremely motivating and challenging features 3 modest: specifically, we do not set out to 4 of a very basic nature, it is impossible to 5 favorable attitudes towards corpus use and 6 d English at primary school. The above data 7 and independent learners. The lessons have 8 es of their connectors to the other groups. 9 ronunciation data in Table 3 contributes to 10 dents said that they probably would use it 11 to write. Though the present study does not 12 t about this finding we ran a t-test, which 13 gical sense of authenticity for motivation, 14 s at any time in a self-access environment, 15 ts he went on to the next task, which again 16 easy nor straightforward. This pilot study	provided that they are evaluated in their provided that enough guidance and sup prove that DDL is better than other meth prove that the students could not have p proved that the introduction of corpora in proved that it was a mixed ability class. proved that the teaching materials based provided that they used them to check proving that both the C and E groups ha (providing that it works), but as mentione prove that this will lead to better writing, proved that all tasks showed a significan provided that students had a secure jump provided that instructional sheets explaini proves that too large a corpus would not b proves that, when the translation is availab
--	---

- 31 Though the corpus is not lemmatised, it is easy to search for all occurrences of [*prove*] by using the asterisk as a wild card (*prov**), and adding *that* to reduce the noise, although some final manual editing will still be necessary to delete lines 10 to 15 (*provided/providing that*). Ten occurrences of [*prove*] *that* is not much in a corpus of 110 research articles, and can be reduced further, as lines 1 to 3 contain negatives (*we do not set out to prove that; it is impossible to prove that; the present study does not prove that*). Further examination in a wider context (FILE VIEW) can exclude a few more occurrences; users who know the corpus well might attribute others to a typical usage by novice researchers or non-native writers of English.

5. Empirical research findings

- 32 The predictable question is whether learners take to the tools and use them successfully. Personal experience with a variety of French L1 learners of English suggests that many of them do. The BYU corpora were used to prepare printed materials for student engineers and architects in their first and second year of higher education respectively (Boulton 2009, 2010a). Students in both contexts managed to use them at least as well as dictionaries under experimental conditions to work out and apply rules of use, even without training. The architecture students were later found to prefer hands-on concordancing, successfully navigating detailed instructions for understanding problem language items over the course of a semester (Boulton 2012c). In distance degree programmes with students majoring in English, third-year undergraduates used the monolingual BYU corpora for translation into the L2 (Boulton 2012d); working on non-literary texts covering specialised themes, their own reported uses during online exams revealed surprising sophistication in their queries and thinking, and led to many appropriate translations. First-year distance master's students following an option in corpus linguistics were required to build their own corpus on a topic related to their personal interests (Boulton 2011b); the vast majority completed satisfactory and sometimes exceptional reports on a wide variety of topics. The uptake is variable even among these English majors, since most are returning to higher education after several years; many are uncomfortable with technology, and others reach master's level without ever having studied linguistics at all. The course requires considerable autonomy from them, exacerbated by the distance degree format. Nonetheless, despite reporting initial trepidation in appropriating the tools, many subsequently elect to use them in their other courses in literature or cultural studies. A recent anecdote may illustrate such spontaneous uses. Early in the semester, an assignment for an entirely separate research methodology course asked them to look at all occurrences of *research* in a specific research article and comment on what they found. On their own initiative, nearly a fifth of the students who submitted the assignment opened the paper in AntConc, which they had been introduced to only very shortly before. These students easily spotted patterns of use which remained invisible to the others, who largely continued to use it as a countable synonym for *study*. Such *ad hoc* and impromptu use shows not only that dedicated corpus tools are accessible to these students, but also that they promote noticing of patterns which otherwise go unrecognised. Charles (2014) is so far the only attempt to look at long-term uptake of corpus use, finding that 70% of her graduate students continued to use their corpus (38% regularly) to help with academic writing.
- 33 Much published research has been interested in learners' receptivity to DDL, typically collected via questionnaires or interviews. The reactions in most cases are positive, often enthusiastically so. Interestingly, Varley (2009) found that enthusiasm was particularly high among students from non-linguistic disciplines who had specific motivations, and who had been less successful with traditional teaching. Nonetheless, it is important not to gloss over possible problems that have been reported. Among those frequently cited are technical issues, though increasingly user-friendly software and interfaces opens up DDL possibilities even among high-school students (Geist & Hahn 2012) and lower-level undergraduates (Boulton 2012c). Learners in a range of situations can use the tools successfully, though not always to maximum effect. Pérez-Paredes et

al. (2012), for example, found the most fruitful results included Google searches alongside corpus queries; however, very little instruction was provided in that particular study. Kennedy and Miceli (2001) are among those to provide detailed analysis of the procedures learners actually use, finding difficulties in the four stages of: formulating the question; devising a search strategy; observing the examples found and selecting relevant ones; drawing conclusions. Sun (2003) identified four factors influencing successful corpus use: prior knowledge of the language point (learners used the corpus successfully to confirm intuitions, but had more difficulty in exploring new points); cognitive skills (comparing, grouping, differentiating, inferring); teacher intervention; and concordancer skills. Most such studies are short term, interested in what students do when they first encounter corpora, and thus provide little or no training; the usual suggestion is for more teacher guidance. This of course depends on how DDL is implemented: paper-based materials or carefully sequenced and scaffolded activities can make for a gentler introduction as learners proceed from “soft” to “hard” DDL where appropriate (e.g. Gabrielatos 2005). That said, software is now a part of learners’ everyday lives, and the concept of searching and making sense of computer-generated responses is no longer as novel or intimidating as it was when DDL first appeared. Some report difficulties coping with truncated concordance lines; sentence formats can be obtained from many concordancers, but generally learners quickly find that the KWIC format makes patterns more visible. Authentic language may be a problem, but using texts relevant to the learners (in their disciplines, chosen by them, or maybe even of their own productions) can make the texts and contents more accessible and enhance a sense of ownership (Charles 2012, 2014). The large numbers of occurrences can lead to a sense of drowning in data, with time-consuming and mechanical sorting sometimes seen as frustrating. Cobb (1999), on the other hand, considers this as an essential component of constructivist learning: while learners may not understand why the teacher does not simply give them the answer, it is the *process* of finding out for themselves that is essential for learning. In other words, an element of difficulty can be a good thing, as long as unnecessary cognitive load is reduced. In the end, however, it must be acknowledged that DDL may not be appropriate for all learners depending on their profiles, preferences, styles, aptitudes, cultures and learning histories. It certainly should not be used as a default resource where other, simpler tools such as dictionaries may be more efficient and relevant (Frankenberg-Garcia 2014). Ultimately, the teacher should decide when and how to introduce corpora appropriately for their students, who will then be in a position to decide when and how to use them for their own purposes.

- 34 In terms of research as a whole, Boulton (2012a) reviewed twenty studies that attempt empirical evaluation of some aspect of corpus use in ESP, finding that learners even with relatively limited levels of proficiency and linguistic sophistication can appreciate them and derive benefit from them in a variety of conditions – on paper or hands-on, as a learning aid or reference tool, etc. In a meta-analysis of 21 DDL studies (not just ESP), Cobb and Boulton (2015) found a mean gain (pre/post-test) of $d=1.68$, and a mean difference (control/experimental group) of $d=1.04$; these figures represent large effect sizes according to Plonsky and Oswald’s (2014) meticulous survey of previous meta-analyses in language teaching and learning as a whole. A more rigorous and inclusive meta-analysis by the same authors (Boulton & Cobb in preparation) identified over 200 empirical studies of DDL, the majority of which involved learners majoring in non-linguistic subjects and/or needing English for specific purposes, with only slightly

lower effect sizes. The papers do cite a number of potential problems (the amount of training required for autonomous use, bewilderment faced with truncated concordance lines, etc.), and the approach may simply not appeal to all individual learners. For any particular group, the only way to find out is to try.

Conclusion

- 35 In a paper such as this we cannot hope to do more than scratch the surface of the possibilities of corpus use; for further reading, excellent general works include Bennett (2010) and Reppen (2010); for ESP and EAP in particular, see Flowerdew (2015) and Gavioli (2005) respectively. Lee and Swales (2006) also report ideas on a corpus-based EAP course for post-graduate students. Our aim here has been to provide some genuine and concrete examples of some of the ways in which a corpus-linguistic mindset can help learners with differing levels of language proficiency and sophistication to approach ESP. DDL draws on solid theoretical foundations and has been found generally to be appealing and useful in terms of learning outcomes. It can be applied even with familiar, everyday tools; this may be enough for some, while others may go further and use existing corpora or even create their own, depending on the investment they are prepared to make, which will in turn depend on their specific or long-term needs and motivations. In targeting specific varieties of English, a corpus approach – for decision-makers, designers, teachers and learners – offers a significant complement to the quirks and failings of intuition alone (Sinclair 2003).

BIBLIOGRAPHY

- ACKERMANN, Kirsten & Yu-Hua CHEN. 2013. "Developing the academic collocation list (ACL): A corpus-driven and expert-judged approach". *Journal of English for Academic Purposes* 12, 235–247.
- ASTON, Guy. 1997. "Small and large corpora in language learning". In LEWANDOWSKA-TOMASZCZYK, B. & J. MELIA (Eds.), *Practical Applications in Language Corpora*. Lodz: Lodz University Press, 51–62.
- ASTON, Guy. 2001. "Learning with corpora: An overview". In ASTON, G. (Ed.), *Learning with Corpora*. Houston: Athelstan, 7–45.
- BARRETT, Louise, Robin DUNBAR & John LYCETT. 2002. *Human Evolutionary Psychology*. Basingstoke: Palgrave.
- BECKNER, Clay, Richard BLYTHE, Joan BYBEE, Morten H. CHRISTIANSEN, William CROFT, Nick C. ELLIS, John HOLLAND, Jinyun KE, Diane LARSEN-FREEMAN & Tom SCHOENEMANN (The 'Five Graces Group'). 2009. "Language is a complex adaptive system: Position paper". *Language Learning* 59 (supplement), 1–26.
- BENNETT, Gena R. 2010. *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers*. Michigan: University of Michigan Press.

- BOULTON, Alex. 2009. "Testing the limits of data-driven learning: Language proficiency and training". *ReCALL* 21/1, 37–51.
- BOULTON, Alex. 2010a. "Data-driven learning: Taking the computer out of the equation". *Language Learning* 60/3, 534–572.
- BOULTON, Alex. 2010b. "Language awareness and medium-term benefits of corpus consultation". In GIMENO Sanz, A. (Ed.), *New Trends in Computer-Assisted Language Learning: Working Together*. Madrid: Macmillan ELT, 39–46.
- BOULTON, Alex. 2011a. "Data-driven learning: The perpetual enigma". In GOŹDŹ-ROSZKOWSKI, S. (Ed.), *Explorations across Languages and Corpora*. Frankfurt: Peter Lang, 563–580.
- BOULTON, Alex. 2011b. "Bringing corpora to the masses: Free and easy tools for interdisciplinary language studies". In KÜBLER, N. (Ed.), *Corpora, Language, Teaching, and Resources: From Theory to Practice*. Bern: Peter Lang, 69–96.
- BOULTON, Alex. 2012a. "Corpus consultation for ESP: A review of empirical research". In BOULTON, A., S. CARTER-THOMAS & E. ROWLEY-JOLIVET (Eds.), *Corpus-Informed Research and Learning in ESP: Issues and Applications*. Amsterdam: John Benjamins, 263–293.
- BOULTON, Alex. 2012b. "Computer corpora in language learning: DST approaches to research". *Mélanges Crapel* 33, 79–91.
- BOULTON, Alex. 2012c. "Hands-on/hands-off: Alternative approaches to data-driven learning". In THOMAS, J. & A. BOULTON (Eds.), *Input, Process and Product: Developments in Teaching and Language Corpora*. Brno: Masaryk University Press, 152–168.
- BOULTON, Alex. 2012d. "Beyond concordancing: Multiple affordances of corpora in university language degrees". *Languages, Cultures and Virtual Communities. Elsevier Procedia: Social and Behavioral Sciences* 34, 33–38.
- BOULTON, Alex. 2015. "Applying data-driven learning to the web". In LEŃKO-SZYMAŃSKA, A. & A. BOULTON (Eds.), *Multiple Affordances of Language Corpora for Data-driven Learning*. Amsterdam: John Benjamins, 267–295.
- BOULTON, Alex, Shirley CARTER-THOMAS & Elizabeth ROWLEY-JOLIVET (Eds.). 2012. *Corpus-Informed Research and Learning in ESP: Issues and Applications*. Amsterdam: John Benjamins.
- BOULTON, Alex & Thomas COBB. In preparation. "Corpus use in language learning: A meta-analysis".
- BOULTON, Alex & Sylvie DE COCK. Forthcoming. "Dictionaries as aids for language learning". In HANKS, P. & G.-M. DE SCHRIVER (Eds.), *International Handbook of Lexis and Lexicography*. New York: Springer.
- BOULTON, Alex & Henry TYNE. 2014. *Des documents authentiques aux corpus: démarches pour l'apprentissage des langues*. Paris: Didier.
- BYRD, Pat & Averil COXHEAD. 2010. "On the other hand: Lexical bundles in academic writing and in the teaching of EAP". *University of Sydney Papers in TESOL* 5, 31–64.
- CHARLES, Maggie. 2012. "'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus-building". *English for Specific Purposes* 31, 93–102.
- CHARLES, Maggie. 2014. "Getting the corpus habit: EAP students' long-term use of personal corpora". *English for Specific Purposes* 35, 30–40.

- COBB, Tom. 1999. "Applying constructivism: A test for the learner as scientist". *Educational Technology Research & Development* 47/3, 15–31.
- COBB, Thomas & Alex BOULTON. 2015. "Classroom applications of corpus analysis". In BIBER, D. & R. REPPEN (Eds.), *Cambridge Handbook of Corpus Linguistics*. Cambridge: Cambridge University Press, 478–497.
- CONROY, Mark A. 2010. "Internet tools for language learning: University students taking control of their writing". *Australasian Journal of Educational Technology* 26/6, 861–882.
- COXHEAD, Averil. 2000. "A new academic word list". *TESOL Quarterly* 34/2, 213–238.
- DUDENEY, Gavin. 2000. *The Internet and the Language Classroom*. Cambridge: Cambridge University Press.
- FLOWERDEW, Lynne. 2015. "Corpus-based research and pedagogy in EAP: From lexis to genre". *Language Teaching* 48/1, 99–116.
- FRANKENBERG-GARCIA, Ana. 2014. "How language learners can benefit from corpora, or not". *Recherches en Didactique des Langues et des Cultures* 11/1, 93–110.
- GABRIELATOS, Costas. 2005. "Corpora and language teaching: Just a fling or wedding bells?" *Teaching English as a Second Language – Electronic Journal* 8/4, 1–35.
- GARDNER, Dee & Mark DAVIES. 2014. "A new academic vocabulary list". *Applied Linguistics* 35/3, 305–327.
- GASKELL, Delian & Thomas COBB. 2004. "Can learners use concordance feedback for writing errors?". *System* 32/3, 301–319.
- GAVIOLI, Laura. 2005. *Exploring Corpora for ESP Learning*. Amsterdam: John Benjamins.
- GEIST, Monika & Angela HAHN. 2012. "Using a corpus for written production: A classroom study". In THOMAS, J. & A. BOULTON (Eds.), *Input, Process and Product: Developments in Teaching and Language Corpora*. Brno: Masaryk University Press, 124–136.
- GELUSO, Joe. 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.
- GILQUIN, Gaëtanelle & Stephan Th. GRIES. 2009. "Corpora and experimental methods: A state-of-the-art review". *Corpus Linguistics and Linguistic Theory* 5/1, 1–26.
- GLEDHILL, Chris. 2000. "The discourse function of collocation in research article introductions". *English for Specific Purposes* 19/1, 115–135.
- HANKS, Patrick. 2013. *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: MIT Press.
- HASSELGREN, Angela. 1994. "Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary". *International Journal of Applied Linguistics* 4/2, 237–258.
- HOEY, Michael. 2005. *Lexical Priming: A New Theory of Words and Language*. London: Routledge.
- JOHNS, Tim. 1988. "Whence and whither classroom concordancing?". In BONGAERTS, P., P. DE HAAN, S. LOBBE & H. WEKKER (Eds.), *Computer Applications in Language Learning*. Dordrecht: Foris, 9–27.
- JOHNS, Tim. 1990. "From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning". *CALL Austria* 10, 14–34.

- JOHNS, Tim. 1991. "Should you be persuaded: Two samples of data-driven learning materials". In JOHNS, T. & P. KING (Eds.), *Classroom Concordancing. English Language Research Journal* 4, 1–16.
- KENNEDY, Claire & Tiziana MICELI. 2001. "An evaluation of intermediate students' approaches to corpus investigation". *Language Learning & Technology* 5/3, 77–90.
- KIRSCHNER, Paul A., John SWELLER & Richard E. CLARK. 2006. "Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching". *Educational Psychologist* 41/2, 75–86.
- LARSEN-FREEMAN, Diane & Lynne CAMERON. 2008. *Complex Systems and Applied Linguistics*. Oxford: Oxford University Press.
- LEE, David & John 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, 56–75.
- LEŃKO-SZYMAŃSKA, Agnieszka & Alex BOULTON (Eds.). 2015. *Multiple Affordances of Language Corpora for Data-driven Learning*. Amsterdam: John Benjamins.
- MANIEZ, François. 2012. "A corpus-based study of adjectival vs nominal modification in medical English". In BOULTON, A., S. CARTER-THOMAS & E. ROWLEY-JOLIVET (Eds.), *Corpus-informed Research and Learning in ESP: Issues and Applications*. Amsterdam: John Benjamins, 85–103.
- MCCARTHY, Michael. 2004. *Touchstone: From Corpus to Coursebook*. Cambridge: Cambridge University Press.
- MCCARTHY, Michael. 2008. "Accessing and interpreting corpus information in the teacher education context". *Language Teaching* 41/4, 563–574.
- MILLAR, Neil. 2011. "The processing of malformed formulaic language". *Applied Linguistics* 32/2, 129–148.
- NESE, Hilary. 2000. *The Use and Abuse of EFL Dictionaries*. Tübingen: Max Niemeyer.
- NESE, Hilary. 2015. "ESP corpus construction: a plea for a needs-driven approach". *ASp* 68, 7–23.
- PARK, Kwanghyun & Celeste KINGINGER. 2010. "Writing/thinking in real time: Digital video and corpus query analysis". *Language Learning & Technology* 14/3, 31–50.
- PÉREZ-PAREDES, Pascual, María SÁNCHEZ-TORNEL & Jose 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.
- PLONSKY, Luke & Frederick L. OSWALD. 2014. "How big is 'big'? Interpreting effect sizes in L2 research". *Language Learning* 64/4, 878–912.
- REPPEN, Randi. 2010. *Using Corpora in the Classroom*. Cambridge: Cambridge University Press.
- SCHMIDT, Richard W. 1990. "The role of consciousness in second language learning". *Applied Linguistics* 11/2, 129–158.
- SCHMITT, Norbert, Tom COBB, Marlise HORST & Diane SCHMITT. 2016. "How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007)". *Language Teaching*. doi:10.1017/S0261444815000075.
- SCOTT, Mike & Christopher TRIBBLE. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.

SIMPSON-VLACH, Rita & Nick C. ELLIS. 2010. "An academic formulas list: New methods in phraseology research". *Applied Linguistics* 31/4, 487–512.

SINCLAIR, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

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

SMITH, Simon. 2011. "Learner construction of corpora for general English in Taiwan". *Computer Assisted Language Learning* 24/4, 291–316.

SUN, Yu-Chih. 2003. "Learning process, strategies and web-based concordancers: A case-study". *British Journal of Educational Technology* 34/5, 601–613.

SUN, Yu-Chih. 2007. "Learner perceptions of a concordancing tool for academic writing". *Computer Assisted Language Learning* 20/4, 323–343.

SWAIN, Merrill. 2006. "Languaging, agency and collaboration in advanced second language proficiency". In BYRNES, H. (Ed.), *Advanced Language Learning: The Contribution of Halliday and Vygotsky*. London: Continuum, 95–108.

TAYLOR, John R. 2012. *The Mental Corpus: How Language is Represented in the Mind*. Oxford: Oxford University Press.

THOMPSON, Paul. 2006. "Assessing the contribution of corpora to EAP practice". In KANTARIDOU, Z., I. PAPADOPOULOU & I. MAHILI (Eds.), *Motivation in Learning Language for Specific and Academic Purposes*. Macedonia: University of Macedonia (n.p.).

TOMASELLO, Michael. 2005. *Constructing a Language: A Usage-based Theory of Language Acquisition*. Harvard: Harvard University Press.

VARLEY, Steve. 2009. "'I'll just look that up in the concordancer': Integrating corpus consultation into the language learning environment". *Computer Assisted Language Learning* 22/2, 133–152.

WORLOCK POPE, Caty (Ed.). 2010. *The Bootcamp Discourse and Beyond*. *International Journal of Corpus Linguistics* 15.

ZAHAR, Rick, Tom COBB & Nina SPADA. 2001. "Acquiring vocabulary through reading: Effects of frequency and contextual richness". *Canadian Modern Language Review* 57/3, 541–572.

Online corpora and tools

AntConc <<http://www.laurenceanthony.net/software/antconc>>

BootCaT <<http://bootcat.sslmit.unibo.it>>

Brigham Young University corpora (BYU) <<http://corpus.byu.edu>>

British Law Report Corpus (BLaRC) <<http://www.lex tutor.ca/conc/eng>>

Business Letters Corpus <<http://www.someya-net.com/concordancer>>

LexTutor (the Compleat Lexical Tutor) <<http://www.lex tutor.ca>>

Linguee <<http://linguee.fr>>

Michigan Corpus of Academic Spoken English (MICASE) <<http://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase>>

Michigan Corpus of Upper-level Student Papers (MICUSP) <<http://micusp.elicorpora.info>>

Oxford Advanced Learner's Dictionary of English (OALD) <<http://www.oxfordlearnersdictionaries.com/>>

University of Oxford Text Archive <<http://ota.ox.ac.uk/>>

WebCorp <<http://www.webcorp.org.uk>>

NOTES

1. Online tools and websites are itemised at the end of the reference list.

ABSTRACTS

A major goal in English for Specific Purposes, however defined, is to provide fine-grained analyses of targeted varieties of English. Corpus linguistics has much to contribute, but still no generic description will be able to cater for all needs and all questions. This paper explores some suggestions for introducing corpus tools and techniques directly to the end-users – learners and teachers – so that they can examine how language is genuinely used in contexts relevant to them. The approach builds on common pedagogical underpinnings (e.g. an inductive approach to authentic documents) and existing practices (e.g. web searches), thus enabling better integration as an everyday activity for different types of users. Beyond web searches, general-purpose corpora can often be queried for specific purposes; users with more clearly-defined or long-term needs may go further and build their own corpus. The paper provides examples of these various uses, and concludes with a brief discussion of what research findings tell us.

L'un des buts de la recherche en anglais de spécialité, quelle que soit la définition que l'on en donne, est de fournir des analyses fines de différentes variétés de la langue. La linguistique de corpus offre de nombreuses possibilités à ce domaine, même si aucune description clés en main ne peut parvenir à répondre à tous les besoins ou à toutes les questions. Cet article suggère quelques propositions destinées à aider l'utilisateur final – apprenant ou enseignant – à mettre en œuvre des outils et des techniques de corpus afin d'observer, en fonction de ses besoins, des usages réels de la langue dans des contextes pertinents. L'approche est articulée avec des approches pédagogiques courantes (par exemple, l'étude exploratoire de documents authentiques) et de pratiques fréquentes (par exemple l'utilisation de moteurs de recherche sur Internet), ce qui permet à différents types d'utilisateurs de mieux intégrer ces outils et ces techniques dans leurs activités quotidiennes. Au-delà des recherches sur Internet, des corpus génériques en ligne peuvent être exploités à des fins spécifiques ; les utilisateurs qui ont des besoins pérennes très ciblés peuvent aller plus loin en construisant leur propre corpus. L'article fournit des exemples de ces divers usages, et conclut par une brève mise en perspective des résultats de la recherche dans ce domaine.

INDEX

Keywords: corpus, data-driven learning, discovery learning, Google

Mots-clés: apprentissage sur corpus, corpus, découverte, Google

AUTHOR

ALEX BOULTON

Alex Boulton is Professor of English and Applied Linguistics at the University of Lorraine. His main research interests broadly relate to English and applied linguistics, and especially corpus linguistics and potential 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: AFLA (vice-president), GERAS, EuroCALL and TaLC; as well as journals such as *ReCALL* (co-editor), *Alsic*, *ASp*, *Eurocall Review*, *IJCALLT*, *JALT-CALL Journal*, *Al-Lisaniyyat*, and *Mélanges Crapel*. He is currently head of the Atilf research group Didactique des langues et sociolinguistique (Crapel), and assistant director of the first UFR Lansad in France (faculty/department of languages for specialists of other disciplines). alex.boulton@univ-lorraine.fr