What has a corpus ever done for you?

Julie Moore assesses the impact of corpora on ELT.

Earlier this year, the University of Birmingham published a list of its top 10 'research heroes' (see Links and references). Nestled in amongst eminent particle physicists, geneticists and Nobel prize-winning chemists was COBUILD, not an individual academic, but a collaborative project originally set up in 1980 between the University and Collins publishers (COBUILD stands for Collins Birmingham University International Language Database). This unique and ground-breaking project, led by Professor John Sinclair, was the first to use large-scale corpus technology to systematically analyse language usage and as a result 'transformed the study of English across the globe'.

That sounds like a big claim, but I don't think it's an overstatement to say that the COBUILD team's approach to understanding how language is used in the real world has revolutionised the way we understand the English language. It's also revolutionised the way we approach compiling not just dictionaries and other reference sources, but also language teaching materials generally and in turn, it has radically influenced the whole field of ELT. No longer is language teaching based in the intuitions of a few individuals and often outdated ideas about what constitutes 'correct usage'; instead what we choose to teach and the patterns of language we now emphasise come from corpus research that helps us mirror in the classroom how the English language is used in the real world.

Nowhere is that influence more evident than in learner's dictionaries. Since the first COBUILD dictionary was published in 1987, the use of corpora as the basis for compiling learner's dictionaries has spread throughout dictionary publishing and now all the major publishers hold their own multi-million word corpora. A modern lexicographer, like myself, wouldn't dream of doing anything without starting from the corpus evidence. Contrary to popular belief, we aren't the fount of all linguistic knowledge and neither do we just 'make it up'. Whether we want to divide up the different senses of a word, determine its most common collocations or find a typical example of usage, it's the corpus we turn to.

That original COBUILD corpus, now called the *Collins Corpus*, has grown and developed over the years; advances in computing power have meant that corpora have been able to grow enormously in size, so that their coverage and their power to represent the language as a whole have improved immeasurably. The dictionaries which the COBUILD project gave rise to have also changed and developed, but they still remain true to their corpus roots, continuing many of the principles established by that original COBUILD team.

That's all great and all very interesting to academics and lexicographers, but as a classroom language teacher, what has a corpus ever done for you? Well, here's my top 10 ways in which corpora have had an impact on ELT, in no particular order:

1 Frequency: words worth learning

One of the key things a corpus can tell us is which words are used most frequently and so, are most useful for students to learn first. The inclusion of words in a learner's dictionary is based on frequency, giving more coverage to the words that will be of most use to the average learner. In COBUILD dictionaries, diamond symbols next to each entry indicate how frequent the word is. This type of information can help learners and their teachers to decide what to focus on. So lower level learners might choose to do more work around high frequency words like *help, watch, hand, care* and *book* (which all have three diamonds), while intermediate learners might focus on the mid-frequency words, like *challenge, doubt, organize* and *profession* (with one or two diamonds).

2 Collocations

The teaching of collocations, words that are commonly used together like *catch a bus* and *best friend*, has now become a central part of ELT. As teachers, we all understand the importance of helping students internalise these common pairings in order to make the language they produce sound more natural and fluent. But it hasn't always been that way. When I started teaching in the early 1990s, there was very little work on collocation in coursebooks beyond the occasional activity around *make* and *do*. It was really when lexicographers were studying corpus lines that they noticed these striking patterns appearing and started highlighting the importance of these key relationships.

Of course, it's still a tricky area for students, because collocations don't always seem to follow logical patterns (why do we say *make an effort,* but *do your best*?). For me, it's an area of language where learner autonomy is key. Students need to know how to check collocations for themselves so that they can improve their fluency and come up with fewer awkward pairings. As a teacher, I'm forever sending my students to the dictionary to check for themselves when they have a query about which words to use together. Being proactive and seeking out an appropriate example or a word partnership box will have more of an impact on their language acquisition than a quick correction from me that goes in one ear and out the other!

3 Phrases and lexical chunks

Producing naturally flowing language isn't just about picking the right two-word pairings either. As researchers explored the patterns that were thrown up by corpora further, they noticed that certain chunks of language kept cropping up together. Sometimes these were the familiar idioms that we all notice so readily, like *once in a blue moon* or *at the end of your tether*, and some were more mundane phrases that we use all the time, like by the way or as soon as possible. Perhaps more interesting are the ones that, as proficient speakers, we don't even notice we're using and might not have previously thought to teach, like *by the look of things or the thing is … .*

There is a risk that students are wont to pick up on a handful of such phrases, learn them by heart and then overuse them – *on the other hand* is a classic case that populates many a student essay! So as with collocations, helping learners to recognise and take on board these lexical chunks needs to be an ongoing process of raising awareness by picking them out when they crop up. And of course, we need to teach the dictionary skills necessary to look up new phrases; where to start looking and how to use cross-references if the phrase isn't at the first word they try.

4 Grammar

And it's not all about vocabulary, corpus research can tell us about the grammar associated with words too. We can learn about the grammatical features of a word itself, for example which adjectives are mainly used attributively, i.e. before a noun. So we talk about *medical problems, medical treatment, medical advice*, etc. but we'd be unlikely to say that 'a problem was medical'. That's shown in the dictionary using a grammar code [ADJ n] (adjective + noun). Corpus evidence allows us to incorporate all kinds of grammatical information into the dictionary; about transitive and intransitive verbs, count and uncount nouns, whether a word is typically followed by an infinitive verb form (*decide to do* [V to-inf]) or a participle form (*enjoy doing* [V v-ing]), whether verbs are typically used in a passive or always followed by a particular preposition. And all this careful analysis is at your fingertips and those of your students with just a little training.

"We all know how difficult it is to come up with natural, typical examples off the top of your head, so I regularly "borrow" examples from the dictionary"

Sometimes a grammar question crops up in class or a student writes something that strikes you as not quite right, but you can't put your finger on why and you don't have a ready answer. My reaction is to try and track down the root of the problem in the dictionary. If I find an answer, I get my students to search for it too, posing the problem and gently pointing them in the right direction, giving hints and clues to help them decode any labels or codes they find. Again, by proactively searching for the answer to a language problem, not only are they more likely to remember the answer when they get to it, but they're also developing skills which will make them more autonomous learners in the future.

5 Authentic examples

When I was learning German at school, I learnt such gems as: '*My leg is broken. Have you got a plaster?*' and '*The rabbit is dead.*' Neither of which, strangely enough, I've ever had the need to use! Thankfully, things have moved on substantially since then and examples in dictionaries and many other teaching materials are now more authentic because they're based on examples of real usage taken from a corpus.

That's great for students, but also for teachers. We all know how difficult it can be to come up with natural, typical examples off the top of your head, so I regularly 'borrow' examples from the dictionary. I use them either just to illustrate the meaning or usage of a particular vocab item, or when I'm putting together a gap-fill or other quick revision activity. The dictionary contains thousands of authentic examples to choose from, so it seems silly not to make use of them.

6 Insights into usage

When you're researching a word using a corpus, you turn up all kinds of fascinating information, some of which doesn't quite fit into any of the standard categories above and can't be slotted neatly into a dictionary entry. That's where usage notes come in: they're a place for all those useful little details; for example, that we mainly use *female* (especially as a noun) in scientific and medical contexts and that we use *dress* as a verb (meaning 'to put on clothes') largely in stories, not in everyday conversation (where we'd say *get dressed* instead). Usage notes can be about register, style or context, they can highlight different regional uses (for example between British and American English) or they might have information about commonly confused words like *lie* and *lay* or *borrow* and *lend*. Usage notes can also be about the thorny issue of 'correct usage', which brings me onto my next point …

7 A descriptive approach

I've lost track of how many conversations I've had explaining that as a language

researcher, my job isn't to tell people what they should and shouldn't say, it's just to pass on useful information about how language is actually used. Modern language research, and by extension language teaching, is about being *descriptive* and *describing* how language is used, not *prescriptive* and setting down rules. That's the joy of corpora: they show us the English language in all its wonderful colour and diversity, warts and all!

That's not to say that we should present that language to learners in a completely neutral, undifferentiated way. It won't help them to learn slang or informal expressions which they then go on to use inappropriately in, say, a job application. One of the major things that corpus research has shown us is the huge difference between spoken and written language. We just don't speak in the same, careful way that we write, and many things that pass completely unnoticed in conversation, would clearly stand out as inappropriate or sloppy in many written contexts. These issues are of course subjective and opinions will differ and often be very strongly held, so dictionaries tend to sit very carefully on the fence. Take, for example, this highly contentious usage: The views are *literally breath-taking*. The COBUILD entry reads 'You can use the word *literally* to emphasize a statement. Some careful speakers of English think that this use is incorrect.'

8 New words

Of course, language is always changing and corpora can help us keep up there too. As corpus data is constantly updated, we can see new trends emerge, not just with the headline new buzzwords like *selfie* or *bestie* (which do make it into learner's dictionaries if they become frequent enough), but also the subtler shifts which often become a part of our linguistic lives more quietly. Take the word *text*, in the 1995 edition of the COBUILD Dictionary, it had only five senses, all nouns to do with 'written material'. Current COBUILD editions contain two extra senses, a noun and a verb, both about messages sent via mobile phones (*I'll text you later*), a usage so ubiquitous now that it would seem odd not to teach it. The same goes for words like *click*, *tablet* and *cloud*.

9 ESP and specialist corpora

With corpora getting ever larger and including a wider range of material (both written and spoken), it's become more feasible to break them down and look at sub-corpora, or sections of a corpus that contain only language from a particular area, such as academic language or more narrowly still, the language used in the field of mechanical engineering. Research in these areas can help us to compile ESP resources such as the *Collins COBUILD Key Words for …* series, which includes vocabulary specific to such fields such as Finance, Accounting, Chemical Engineering or Retail.

"One of the major things that corpus research has shown us is the huge difference between spoken and written language."

10 The digital revolution

Whilst many areas of ELT are still getting to grips with how to make the best use of digital media, dictionaries have been way ahead of the curve. Learner's dictionaries have been about in the form of CD-ROMs, handheld electronic dictionaries and online for many years, and mobile apps are now commonplace. The nature of a dictionary lends itself perfectly to a digital format, with discrete entries that fit neatly on a single screen and lots of functions which so obviously make using a dictionary much easier, like flexible search facilities, the ability to have audio for pronunciation and simple hyperlinks instead of slightly clunky (and often ignored) cross-references.

One of the challenges of the internet age is helping learners to navigate through the mass of reference information available. If you search for a word online, you'll come up with a variety of online dictionaries, some of dubious quality and many aimed primarily at a native-speaker audience who have different priorities from the average language learner. Nowadays, teaching the skills that learners need to access appropriate, good quality reference information also involves an element of digital literacy. As teachers, we need to spend some time linking up the learner's dictionaries on the shelves or desks in the classroom to the gadgets, apps and websites that students access all the time, and pointing them in the direction of what's most useful and appropriate for them.

With such a fantastic wealth of information about language out there though, thanks largely to the work of COBUILD and the corpus researchers that followed them, there's really no excuse for teachers and learners not to make full use of it!

Links and references

'University of Birmingham's "top ten" celebrate global impact': http://www.birmingham.ac.uk/news/latest/2014/05/University-of-Birminghams-top-ten-celebrates-global-impact.aspx

COBUILD free online dictionary: www.collinsdictionary.com/cobuild

COBUILD Advanced Learner's Dictionary (8th edition) (2014): http://bit.ly/COBUILDAdvanced

COBUILD Intermediate Learner's Dictionary (3rd edition) (2014): http://bit.ly/COBUILDIntermediate

Julie Moore is a freelance ELT teacher, writer and lexicographer based in Bristol in the UK. She first became interested in lexicography and corpus research doing her MA at the University of Birmingham and over the past 15 years, she has worked on learner's dictionaries for all the major ELT publishers. She now combines teaching, teacher training, lexicography and other ELT writing and never ceases to be amazed that she can earn a living from playing with words!