

Corpora in ELT¹

Ana Frankenberg-Garcia

Introduction

This chapter begins with an explanatory definition of corpora and covers different types of corpora available for ELT. It then proceeds to explain what corpus software does and how corpora can be used in ELT. It concludes with a discussion of key areas of debate surrounding the use of corpora for language teaching and future directions in this domain. The chapter does not assume any prior knowledge of corpora

What is a corpus?

In very simple terms, a corpus is a collection of texts in electronic format. Although many people use the word ‘corpora’ simply to refer to a body of texts, in the present chapter, the word is being used in the corpus linguistics sense, where those texts must be in a machine-readable format. Thus, texts that are only available in print will need to be digitised before inclusion in a corpus. Likewise, in order to include spoken language in a corpus, speech needs to be recorded and then digitally transcribed into text.

It is important to bear in mind that, unlike electronic libraries, which store texts for their intrinsic value, the texts compiled into a corpus are normally selected so as to be fit for a particular purpose. A corpus that is to be used to teach English academic writing, for example, may include academic essays, journal articles and dissertations in English, but not genres such as fiction and news.

Many scholars also believe that the texts in corpora need to have been produced for authentic communicative goals (e.g. Sinclair, 1991; Tognini-Bonelli, 2001). A corpus for teaching business English, for example, should include business letters and transcripts of business meetings that actually took place rather than letters invented for the purpose of teaching business English or transcripts of staged meetings recorded solely for the purpose of language teaching. This is because it is believed that only “genuine communications of people going about their normal business” (Tognini-Bonelli, 2001: 55) can give us a true picture of the language that is actually used in those circumstances. Studying such texts reveals facts about language that would otherwise remain unnoticed.

Two other key aspects of corpora are their size and representativeness. Common sense dictates that corpora should be large enough to allow one to make useful generalisations about the language represented in a particular corpus. For example, one cannot draw any conclusions about how to write abstracts by analysing only one or two abstracts. However, a small corpus of, say, forty medical journal abstracts can help one detect patterns about the language and organisational structure of this highly specific genre. On the other hand, corpora used for the analysis of general language often need to be quite large, for they must contain a sufficient number of texts which are representative

¹ This is a pre-publication author version of a chapter to appear in Hall, G. (ed.) (2016) *Routledge Encyclopaedia of ELT*.

of a wide range of situations. The ideal size of a corpus will ultimately depend on what the corpus is for. This can range from small, specialised language corpora for teaching ESP, with just a few thousand words (see Gavioli, 2005), to very large reference corpora with millions or even billions of words used in corpus-based lexicography.

The four essential characteristics of corpora described above are neatly summarised by McEnery et al. (2006: 5), for whom a corpus is “a collection of (1) machine-readable (2) authentic texts which is (3) sampled to be (4) representative of a particular language or language variety”.

Different types of corpora

In the same way as there are different types of texts in the world, there are also different types of corpora. Table 1 lists a selection of open-access online corpora that can be used directly by EFL teachers and learners.

The British National Corpus (BNC; see Table 1), for example, was compiled in the early 1990s with the purpose of providing a snapshot of British English; thus it can be said to be a general language corpus. It contains a wide variety of texts, from formal academic writing to transcripts of spoken teenage language. Although it is still widely used in research and teaching, the BNC does not contain words or meanings that are relatively new in English; it reflects British English at a particular point in time. Thus a word like *wireless* will appear in the BNC in its rather old-fashioned sense meaning radio, but not in its current widely used form as a modifier in phrases like *wireless phone* and *wireless network*. With 100 million words, the BNC is also relatively small by today's standards. The dramatic increase of digital texts in the world since the 1990s has made corpora much easier to compile. The corpus underlying SkELL (Sketch Engine for English Language Learning; see Table 1), for example, contains over one billion words gathered from British and American websites (Baisa and Suchomel, 2014), providing a good coverage of everyday, standard, formal and professional English.

As explained above, however, corpora do not have to be huge to be useful. The academic language corpora in Table 1, for example the British Academic Written English corpus (BAWE), are much smaller. Likewise, the Business Letter Corpus (BLC; see Table 1), with just one million words, is a good example of a small, specialised corpus of American and British business letters that can be used in teaching business English.

Corpora consisting of speech also tend to be small, because it takes time to transcribe speech and not all forms of conversation are easy to capture. While parliamentary debates and radio and television talk shows are recorded anyway, in order to collect transcripts of other forms of speech, it is first necessary to ask volunteers to go around recording their own everyday conversations. For this same reason, the spoken component of general language corpora like the BNC or the Corpus of Contemporary American English (COCA; see Table 1) tend to be much smaller than the written one.

Table 1: A selection of open-access corpora with integrated online concordancers that can be used directly by teachers and learners of English

Corpus	Size in words (millions)	Brief Description	URL	Relevance to ELT
British Academic Spoken English Corpus (BASE)	1 M	British university lectures and seminars	https://ca.sketchengine.co.uk/bonito/run.cgi/first_for_m?corpname=preloaded/base	Spoken academic English
Michigan Corpus of Academic Spoken English (MICASE)	1.8 M	Academic speech at the University of Michigan	http://quod.lib.umich.edu/mmicase/	
British Academic Written English Corpus (BAWE)	6.5 M	British university student essays	https://ca.sketchengine.co.uk/bonito/run.cgi/first_for_m?corpname=preloaded/bawe2	Written academic English
Michigan Corpus of Upper-Level Student Papers (MICUSP)	2.6 M	American grade A student papers	http://micusp.elicorpora.info/	
Business Letter Corpus (BLC)	1 M	British and American business letters	http://www.someva-net.com/concordancer/	Business English
British National Corpus (BNC)	100 M	British English from the early nineties	http://corpus.byu.edu/bnc/ (also available from other online concordancers)	General English
The Corpus Of Contemporary American English (COCA)	450 M	American English from 1990 to 2012	http://corpus.byu.edu/cocac/	
Corpus Of Global Web-Based English (GloWbE)	1.9 B	Web pages from different English-speaking countries	http://corpus2.byu.edu/glowbe/	
SKELL	1 B	A interface-com-corpus conceived for ELT based on recent texts gathered from the web.	https://skell.sketchengine.co.uk/run.cgi/skell	
Vienna-Oxford International Corpus of English (VOICE)	1 M	Speakers from different first language backgrounds using ELF	http://www.univie.ac.at/voice/boel/corpus_availability_online	ELF
OPUS, the Open Parallel Corpus	open-ended	A collection of source texts and translations in English and various other languages, with several specialized language subcorpora for ESP	http://opus.lingfil.uu.se/	Use of L1 to teach English

In addition to corpora of speech and writing, there have been recent efforts to compile multimodal corpora, such as the Padova Multimedia English Corpus (Cocetta, 2011). These corpora contain written transcripts of speech aligned with video recordings so that it is possible to study how language and non-linguistic elements such as gesture, facial expressions and gaze are used in conjunction to create meaning.

When referring to a corpus of English, the default is to assume the texts in the corpus were produced by native speakers (this chapter will not deal with the debates surrounding the term ‘native speaker’; however, see Llorca, this volume, for further discussion). However, there are certain types of corpora that focus precisely on the language of non-native speakers. Learner corpora, made with texts produced by learners of English, are compiled to research learner error and second language development. The International Corpus of Learner English (Granger, 2003) is a notable example of this kind of corpus. Likewise, there are also corpora that have been assembled to study English as a lingua franca (ELF; see Seargeant, this volume), like the Vienna-Oxford Corpus of English (VOICE; see Table 1), which comprises transcripts of conversations by users of English as an international language.

In addition to monolingual corpora, there are also parallel corpora consisting of source texts aligned with their translations into another language. These corpora tend to be smaller, specialised language corpora – of parliamentary debates, of fiction, of film subtitles, for example – because they can only include texts belonging to genres that are available in translation. As shall be seen later in this chapter, parallel corpora can be particularly useful in ELT when one wishes to highlight L1 and L2 contrasts.

Finally, while all corpora consist of plain text, some corpora contain extra information attached to the text that is not normally visible to the corpus user. The two most common addons are ‘lemmatisation’ and ‘part-of-speech tagging’. Lemmatisation involves labelling each word in a corpus with its base form, i.e. its lemma. Thus in a lemmatised corpus, ‘hidden’ behind the sentence *I was fifteen minutes late* is the information shown below in brackets:

*I*_[lemma=I] *was*_[lemma=BE] *fifteen*_[lemma=FIFTEEN] *minutes*_[lemma=MINUTE]
*late*_[lemma=late]

Lemmatisation allows one to carry out queries which, in a single action, retrieve all the inflections of a given word. For example, a corpus search for [*lemma=BE*] followed by *late* will retrieve *am late*, *'m late*, *is late*, *'s late*, *are late*, *was late*, *were late*, *been late* and *being late*, which is more practical than looking up each of these separately. This is also a good way of retrieving sentences for gapping exercises for learners to practice the verb *to be*.

Part-of-speech (POS) tagging, in turn, involves labelling each word in a corpus with its POS category. The example above would be tagged as follows:

*I*_[pos=PRONOUN] *was*_[pos=VERB] *fifteen*_[pos=NUMBER] *minutes*_[pos=NOUN]
*late*_[pos=ADJECTIVE]

This kind of tagging allows one to carry out sophisticated queries involving POS categories. For example, a search for *I* [*pos=VERB*] *late* will retrieve *I am late*, *I slept late*, *I arrived late*, *I work late*, and so on. This is a practical way of retrieving sentences that can be transformed into a vocabulary exercise for learners.

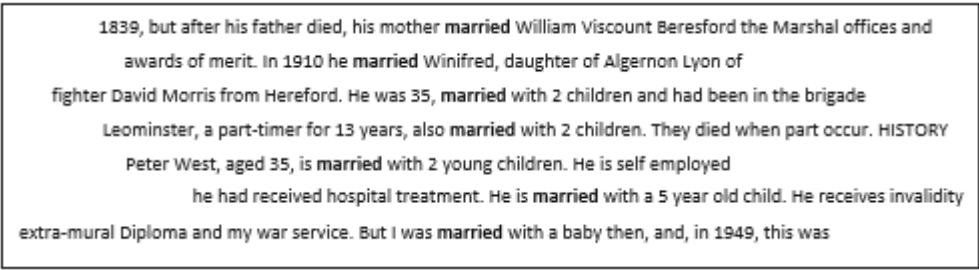
Apart from lemmatisation and POS tagging, it is possible to add all sorts of extra information to a corpus that will help one search the corpus more efficiently. Learner corpora, for example, are tagged for errors so as to facilitate their retrieval and analysis.

Corpus software

Containing many thousands or even billions of words, corpora are used in conjunction with special text-retrieval software known as concordancers. Concordancers enable one to manipulate and interrogate a corpus in a way that is very different from reading texts from start to finish, often providing insights into the language represented by the corpus which are not visible to the naked eye. Concordancers perform three basic types of operations, generating concordances, word lists and collocation statistics. These will be now be explained.

Concordances

Concordance queries work in a way similar to the *find* function of a normal electronic text editor. However, instead of skipping from one occurrence of whatever we look up to another, the concordance option lists all such occurrences together, displaying them vertically along with the context in which they appear, as exemplified by the sample concordances for *married* in Figure 1. This simple key-word-in-context (KWIC) display allows one to scroll down the computer screen and notice various patterns of how *married* is used. The concordances in Figure 1 have been sorted alphabetically one word to the right of *married*, enabling learners to focus on the patterns used after *married with*. This could be useful for those who make errors like *‘*She is married with a Frenchman*’.



1839, but after his father died, his mother married William Viscount Beresford the Marshal offices and awards of merit. In 1910 he married Winifred, daughter of Algernon Lyon of fighter David Morris from Hereford. He was 35, married with 2 children and had been in the brigade Leominster, a part-timer for 13 years, also married with 2 children. They died when part occur. HISTORY Peter West, aged 35, is married with 2 young children. He is self employed he had received hospital treatment. He is married with a 5 year old child. He receives invalidity extra-mural Diploma and my war service. But I was married with a baby then, and, in 1949, this was

Figure 1 Sample KWIC concordances for *married* from the BNC

In addition to the standard KWIC display in Figure 1, which focuses the user’s attention on the right- and left-hand co-text of a word or string of words, many concordancers allow users to switch to a full sentence view instead. This can be particularly important when teaching discourse, as can be seen from the sample concordances in Figure 2, all of which highlight the sentence-final position of the adverb *tomorrow*. Some concordancers also allow users to expand the co-text of each concordance line so as to retrieve more context preceding and following a given concordance.

Concordance queries in parallel corpora will in turn retrieve parallel concordances consisting of aligned source text and translation segments. Figure 3 contains a selection of parallel concordances from the COMPARA corpus (open access at www.linguateca.pt/COMPARA; see also

Frankenberg-Garcia and Santos 2003) which draw attention to the fact that the Portuguese adverb *atualmente* does not translate into its English cognate *actually*. This could clearly be useful for learners when considering lexical ‘false friends’, for example.

My son is turning 30 years old **tomorrow**.
 The trial was adjourned until 10am **tomorrow**.
 The tours are supposedly being indefinitely suspended effective **tomorrow**.
 My summer young artist applications are due **tomorrow**.
 The city board holds its regular monthly meeting **tomorrow** night.
 A small shift today becomes a disaster **tomorrow**. I promise you something less serious **tomorrow**.

Figure 2 Sample full-sentence concordances for *tomorrow* from SkELL

O original pertenceu mais tarde a Luis XIV e atualmente está no museu do Louvre.	The original was later a possession of Louis XIV and hangs now in the Louvre.
-- Que meios de subsistência tem ele atualmente ?	`What are his present means of subsistence?'
Como está ela atualmente ?	How's she these days ?
Atualmente finjo que não ouço.	I tend to ignore it nowadays .
Atualmente existe uma espécie de epidemia de falta de amor-próprio em Inglaterra.	There's something like an epidemic of lack of self-esteem in Britain at the moment .

Figure 3 Sample parallel PT > EN concordances for *atualmente* from COMPARA

Note finally that concordance queries need not focus on single words. Users can also search for conventional strings of words like *if I were you* (Figure 4), and, as explained above, depending on the corpus, it is possible to carry out more sophisticated queries involving lemmas, POS categories and other types of corpus annotation. Figure 5 exemplifies concordances for *a [pos=ADJECTIVE] escape*.

I should stay lying down **if I were you**.
 I would leave **if I were you**.
 I'd move on **if I were you**.
 I should try to forget it **if I were you**. I'd go home **if I were you**.
 I would watch my step **if I were you**.

Figure 4 Sample concordances for *if I were you* from SkELL

I never saw such **a fast escape**.
 They might have had **a miraculous escape**.
 He had **a lucky escape** from execution.
 They shot him during **an alleged escape**.
 It had been **a narrow escape** and I was impressed.
 Write a short story about **a daring escape** attempt.

Figure 5 Sample concordances for *a [pos=ADJECTIVE] escape* from SkELL

Word lists

Word lists are simply lists of all the words in a corpus along with information about their frequency and rank in the corpus. From the perspective of ELT, they can be very useful to help one determine what vocabulary to teach first. According to Zipf's law (Zipf, 1949), the top most frequent words in a language cover a very large proportion of the language as a whole, so if learners are able to understand and use, say, the 3,000 most frequent words, they should in theory be able to get by in most situations. As Cook (1998: 58) stated, however, corpora can only supply us with "information about production but not about reception". Corpus frequencies alone should therefore not be the sole criterion used when selecting what language to teach.

Apart from plain word lists, some concordancers also allow one to extract 'keyword' lists. This is typically done by comparing word frequencies in a specialised language corpus with word frequencies in a general language corpus. The words that are particularly salient in the former will be ranked first, highlighting the peculiarities of the specialised language in question. For example, by comparing verb frequency in the BLC with verb frequency in a corpus of general English, Someya (1999) was able to generate a list of verbs like *thank, enclose, appreciate, request, order, receive, schedule, attach, purchase, discuss* and so on that are particularly significant in business letters. Of course, it was only possible to isolate verbs in this way because the BLC is tagged for POS.

Using the same methodology, it is also possible to extract frequency lists of strings of words in order to identify key phrases in a corpus. Examples of core five-word strings in the Business Letter Corpus are: *thank you very much for, look forward to hearing from, do not hesitate to contact, please let me know if* and so on. Careful scrutiny of such a list can be very useful when it comes to identifying and selecting phrases that are typical of business letters.

Collocations

Proficient language users know instinctively which words go together in a language and which words sound awkward when combined. Some collocations are dictated by logic, like the verb *drink* followed by liquids like *water, beer* and so on, while others are purely arbitrary and often differ from language to language. For example, it is conventional to say *auburn hair*, but people do not say *auburn scarf*, even when the two are exactly the same colour. The concept of collocation (Firth, 1957) pre-dates corpus linguistics and collocation statistics. However, with the emergence of electronic corpora, it is now possible to list collocations in seconds by running automatic statistical calculations that compare the overall frequency of particular words in a corpus with their frequency in the immediate context of another word. This will show how likely it is for the words in question to combine. Imagine a learner trying to think of a verb to follow the noun *situation*. A collocation query would automatically list verbs like *arise, worsen, escalate, deteriorate, exist, warrant, change, improve, affect* and so on, which is more efficient than running a concordance query for *situation* and scrolling down the results until a suitable verb was found. It would be equally simple to run a collocation query in order to list adjectives that collocate with *situation*, like *dire, hopeless, desperate, untenable, tense, emergency, current, financial, stressful, win-win* and so on. This can be extremely useful to help learners expand their vocabulary and write more idiomatically.

Types of concordancers

As explained in the beginning of this section, most concordancers will allow users to run concordance, word list and collocation queries. However, their interfaces vary, and so does the query language associated with them. Some concordancers are more sophisticated than others,

allowing users to run queries that are not possible in simpler software. Users can install proprietary standalone concordancers like WordSmith tools (Scott, 2012) or freeware like AntConc (Anthony, 2014) and use them to interrogate corpus files stored on their personal computers. Alternatively, as previously exemplified in Table 1, there are numerous open-access corpora that can be interrogated remotely via an online interface without any software installation. SkeLL was purposefully conceived for ELT (Baisa and Suchomel, 2014) and is arguably the most user-friendly English corpus-cum-concordancer available today.

Uses of corpora in ELT

Corpora are used to develop various general language tools that have become commonplace in people's lives, including spell checkers, autocorrect options in text editors and web browsers, and even sophisticated machine-translation programs.

In state-of-the-art pedagogical lexicography, corpora are employed to research word use, and this information is collated to select which headwords (i.e. words listed in a dictionary) are important to include in learners' dictionaries, which senses of polysemous words to present first, which words to use in the definitions, and which grammatical properties and collocations of words to draw attention to. Modern learners' dictionaries also provide corpus-based examples that can help learners see how words are used in context and utilise data from learner corpora to draw attention to recurrent errors. For example, the word *information* in the Macmillan English Dictionary online is marked with three stars, meaning it is a very frequent word in English. The entry for *information* also shows common collocates and phrases, such as *get/obtain/collect information, information about/on/regarding, a piece of information, relevant/useful information, further information*, contextualised examples such as *We were able to get the information we needed from the Internet*, and a 'get it right' rubric explaining that *information* 'is never used in the plural or comes after *an* or a number'. This explanation is then exemplified with learner corpus data of what is wrong – **TV helps people to get an important information* – and how to correct it: *TV helps people to get important information*.

Apart from dictionaries, at least in the ELT market, there is a growing body of grammars, coursebooks and even language tests that draw on corpus data to develop their content in a number of different ways. Corpus frequencies may be used to inform what words and phrases to include in a syllabus and to distinguish between language that is typically spoken, written, formal and informal (see Biber et al., 1999, for example). Concordance lines may be incorporated into dialogues and exercises, learner corpus data may be used to identify problematic areas that require special attention, and so on. In the Touchtone Series (McCarthy et al., 2005), for example, learners are informed that 'People say *Sometimes I* seven times more often than *I sometimes*' (p.46). Figure 6 lists a few well-known corpus-based ELT publications.

The publications in Figure 6 contain language that has been selected from raw corpus data and edited by lexicographers, corpus linguists and materials designers. Of course, the amount of language that can be presented in this polished format is limited, simply because language is infinitely bigger and more complex than what can be summarised in a book or any other language learning aid. Language learners (and their teachers) often have questions for which there are no answers or which are not treated in sufficient detail in dictionaries, coursebooks, grammars and other educational publications.

Cambridge Dictionary of American English	Macmillan English Dictionary
Cambridge International Dictionary of English	Macmillan Collocations Dictionary
Cambridge Grammar of English	Natural Grammar (Oxford)
Collins COBUILD English Dictionary for Advanced Learners	Oxford Advanced Learner's Dictionary
Collins COBUILD English Usage	Oxford Collocations Dictionary for Students of English
Collins COBUILD Intermediate English Grammar	Practical English Usage (Oxford)
Longman Dictionary of Common Errors	Touchstone series (Cambridge)
Longman Dictionary of Contemporary English	Vocabulary in Use series (Cambridge)
Longman Grammar of Spoken and Written English	

Figure 6 Examples of corpus-based ELT publications (source: Frankenberg-Garcia, 2014)

Thus, another option is for teachers and learners to use corpora directly. Corpora can provide more language and can disclose solutions to language queries that have not been dealt with in edited language resources, propelling language users to completely new levels of learner autonomy (see Benson, this volume). The direct use of corpora has come to be known as discovery or data-driven learning (DDL). For Johns (1991: 3), the founding father of DDL, “What distinguishes the data-driven learning approach is the attempt to cut out the middleman . . . and give direct access to the data so that the learner can take part in building up his or her own profiles of meanings and uses”.

Language teachers do not normally have time to compile corpora and conduct corpus-based ELT research, but they can resort to ready-made corpora to complement their teaching in two basic ways. They can prepare corpus-based handouts and exercises for their students, and they can teach learners to use corpora on their own. Gabrielatos (2005) has referred to this distinction as the ‘soft’ and the ‘hard’ approach to using corpora in the classroom, while Boulton (2010) prefers to call this the ‘hands-off’ and the ‘hands-on’ approach. An example of the former is given in Frankenberg-Garcia (2012a). A group of Portuguese learners of English did not understand the meaning of *aisle* when they were exposed to the word via a dialogue in their coursebooks. The word appeared in the context of air travel, and its meaning in that sense was briefly explained. For the following lesson, the teacher prepared the exercise in Figure 7 in order to expand and consolidate the learners’ previous one-off contact with the word. As can be seen, the *aisle* exercise enhances the learners’ exposure to the new input by presenting them with concentrated doses of the word in context. The learners were able to figure out that aisles exist not just on aeroplanes (which was the original context in which they had seen the word) but also in places like trains, shops, churches and supermarkets. Additionally, the concordances served to help the learners notice that there is a distinction between aisles and corridors, which does not apply to their native language.

Frankenberg-Garcia (2012a) also gives an example of hands-on corpus consultation during a session in which learners were looking at different ways of ending business letters. One of the students was not happy about *I look forward to hearing from you*. She said her former tutor (a native speaker of English) had told her that the right way of saying this was using the present continuous: *I am looking forward to hearing from you*. The teacher (a non-native speaker) felt both forms were correct, but for reassurance she and the student looked up the strings *look/looking forward to hearing/seeing* in the BLC. The results summarised in Table 2 showed the student, and confirmed to the teacher, that not only it was perfectly acceptable to end a letter with *look forward to hearing/seeing*, but also that it seemed in fact to be more conventional than *looking forward to hearing/seeing*.

At a more advanced level, Charles (2012) has taught non-native PhD students to build their own corpora in their specialised domains to help them research the specialised terminology and phraseology of their fields of study.

A. Read the sentences below and make a list of the sort of place where aisles are found.

B. Does aisle translate into Portuguese always in the same way?

1. The air hostesses inquired what I was making and a man passing in the aisle quite genuinely complimented me on my work.
2. I arrived at Salisbury Cathedral, just as the bride was about to go up the aisle.
3. As she looked around she felt a twinge of sadness that in a carriage where 70 per cent of the commuters were men there were five women forced to stand in the aisle.
4. They looked at the passports and then started to walk down the aisle, pointing their guns at the passengers.
5. He hurried up the aisle of the church.
6. She picked up her suitcase and made her way along the aisle.
7. The layout of the store, with wide aisles, gives customers room to move around.
8. I spend much of my time at the shops; wandering through the aisles, flustered, never knowing what to buy.

Figure 7 Handout with selected BNC concordances for *aisle* (source: Frankenberg-Garcia, 2012a: 40)

To summarise, therefore, corpora have been used by linguists and lexicographers to write dictionaries, grammars and coursebooks and by teachers to prepare materials for their students or to help themselves and their students learn more about a language by consulting corpora directly.

Table 2 Distribution of *looking/look forward to seeing/hearing* in the BLC

Search string	Corpus Frequency
look forward to hearing	212
looking forward to hearing	19
look forward to seeing	156
looking forward to seeing	15

Key areas of debate

Authenticity

One of the most widely debated issues surrounding the use of corpora in ELT is the actual language represented in corpora. As explained earlier, corpora are made of texts taken from real-life communications between people. Corpus-based ELT materials therefore draw on attested language use rather than on language invented for the purpose of teaching. Attested language is often described as ‘authentic’, ‘genuine’ or ‘real’ language, and corpus-based publications have capitalised on this to market their products. For instance, Cambridge’s Touchstone series (McCarthy et al., 2005) claims to teach “English as it’s really used [and] presents natural language in authentic contexts”. Similarly, the motto of Collins COBUILD Dictionary online is “supporting learners with authentic English since 1987”, and the Longman Dictionary of Contemporary English online advertises that “155,000 natural examples bring English to life”.

Widdowson (2000), however, points out that corpus data cannot be regarded as authentic once it has been uprooted from its original context. In other words, a text can only be regarded as authentic by those who use it for natural communicative purposes; when reused in a classroom for the

purpose of teaching and learning a language, strictly speaking, it can no longer be said to be authentic.

Possibly more important than discussing the term authenticity, however, is the fact that because corpora are made of genuine communications, they may include errors, taboo words, sensitive topics that are not appropriate for classroom use, and, in particular, language that is simply too difficult for learners. The argument here runs that unreal or scripted language is more accessible for learners and therefore more pedagogically appropriate. For Widdowson (1998: 714–715),

the whole point of language learning tasks is that they are specifically contrived for learning. They do not have to replicate or even simulate what goes on in normal uses of language. Indeed, the more they seem to do so, the less effective they are likely to be.

Yet the point of using corpora in ELT is not to impinge raw corpus data on learners, but, as seen in the examples given in the previous section, to process this data so as to extract from it facts about language that are often unavailable elsewhere, promoting learner autonomy. If preparing hands-off corpus based hand-outs and exercises beforehand, teachers need to use their common sense so as to edit out corpus data that is unsuitable for teaching purposes. This cannot of course be done when learners use corpora hands-on, in which case teachers must be prepared to deal with exposing learners to raw corpus data. While this is very far from the idealised language that we normally see in ELT materials, it must be recognised this is also language people in the streets actually use. It could therefore be argued that corpora provide a golden opportunity for language learners to be able to get in touch with the language people use outside the classroom in the sheltering presence of a teacher.

Apart from the fact that attested uses of language can better prepare learners to communicate effectively and competently in real life outside the classroom, concordance data is often more interesting and thought-provoking than the sometimes insipid and contrived language used in scripted textbook dialogues and exercises. Römer (2004: 153) stresses this point by comparing particularly unrealistic sentences from a traditional German textbook dialogue like *'Where are the girls? Are they packing? Yes, they are.'* with spoken data from the BNC like *'What's happening, does anybody know? Are you listening to me?'*

There are however many very good course materials that are not corpus-based. Moreover, simplifying and adapting a language so as to make it more accessible to language learners is not necessarily a bad idea. Parents do this instinctively when speaking to babies and toddlers, and native speakers tend to do this when communicating with non-native speakers; whatever their pitfalls, coursebooks with invented sentences and scripted dialogues have been useful in helping people to learn foreign languages for many generations.

Although there are a number of studies on how teachers and learners react to corpora (see Boulton and Pérez-Paredes, 2014, for example), the fact is that further research is needed in order for us to come to a better understanding of how presenting learners with attested instances of language use from corpora compares with the idealised language that we often see in more traditional ELT textbooks.

Corpora for ELT

Another important issue is choosing a suitable corpus for classroom use. Traditional ELT materials tend to have been written, or at least edited, by native speakers of English, and corpus-based ELT publications on the market follow suit by using large native-speaker corpora (albeit often with insights from non-native, learner corpus data).

Since the language of these corpora may be too difficult or unsuitable for learners, some scholars believe in creating corpora purposefully designed for language teaching rather than using general language corpora like the BNC or COCA (see Table 1). The SACODEYL project (Widmann et al., 2011), for example, provides online access to very small pedagogic corpora in seven European languages consisting of video-recorded interviews with 13- to 17-year-old teenagers. This is an example of a corpus that was purposefully compiled for teaching young learners.

With the emergence of ELF, it is also important to consider how relevant corpora like the VOICE corpus in Table 1 might be for ELT (Seidlhofer, 2004). The question of whether ELF corpora are purely for linguistic research or whether they can or even should be used to inform ELT teaching is part of the ongoing debate on recognising L2 English users as speakers in their own right rather than as 'failed' native speakers.

Another issue related to the type of corpora used for language teaching is in what situations it is legitimate to use parallel corpora, since parallel concordances will inevitably put learners in contact with (a) L1 and L2 contrasts and (b) translated language. Frankenberg-Garcia (2005, 2007) argues that parallel corpora can be especially useful in teaching monolingual classrooms in a number of situations where comparing L1 and L2 is beneficial (see Svalberg for discussion of similar ideas in relation to language awareness, and Kerr for discussion of use of learners' own-language in the ELT classroom, both this volume)

Direct uses of corpora by teachers and learners

A further key area of debate surrounding the use of corpora in ELT is teachers' and learners' reactions to direct uses of corpora in the classroom. Even though the number of corpora that can be used by anyone with access to computers and the Internet has taken a giant leap over the past few years, there are still very few teachers, let alone learners, who feel comfortable using corpora directly. One of the problems is that most corpora were compiled for research rather than for teaching purposes, and the use of most concordancers is not very intuitive. There are a number of studies (for example, Kennedy and Miceli, 2001; Frankenberg-Garcia, 2012b) that show that corpus skills do not come naturally and need to be taught.

However, mastering the basics of corpora (i.e. how to select an appropriate corpus, how to work with concordances, word lists and collocations, and, most importantly, how to interpret corpus data) is not the only difficulty. The next major problem is how and when to transpose this expertise to the classroom. The fact that a ready-made ELT publication or a custom-made exercise or activity prepared by a teacher is corpus-based does necessarily mean that it is good. As with any other ELT material, corpus-based teaching aids must be relevant, useful and accessible to the particular group of learners they were designed for. Likewise, there are hundreds of ways in which learners can explore corpora on their own, but first they must develop a sense for what queries might be useful to them and understand what to do with the data retrieved. Unfortunately, there seem to be quite a number of corpus-based activities exemplified in the literature which have more to do with linguists' interest in language research than with language learners' actual needs. Language learners (and

their teachers) cannot be expected to compile corpora and be captivated by analysing corpus data just because this is fascinating to linguists.

A final problem presented by data-driven learning is how to fit it in with the rest of the teaching curriculum. The few teachers who are using corpora today seem to be teachers who do research in corpus linguistics and work in an environment – mostly universities – where they have a great deal of autonomy regarding what and how they teach. However, this is not the case in the majority of ELT scenarios. As discussed in Frankenberg-Garcia (2012b), most teachers are not researchers. They normally have a syllabus to follow and do not have much time for devising corpus-based activities, let alone compiling corpora. Moreover, language lessons do not normally take place in computer labs that enable hands-on access to corpora.

Future directions

While the tendency is for there to be more and more corpus-informed ELT publications on the market, the direct use of corpora by teachers and learners is something that has yet to be addressed. The need for training pre-service and in-service teachers to use corpora is acknowledged by several scholars (for example, Mukherjee, 2004; Römer, 2009; Frankenberg-Garcia, 2012b; Lenko, 2014). It is only when language teachers have learnt how to use corpora that their expertise can trickle down to benefit language learners as well. Even teachers who do not have the time or are not willing to use corpora with their students can use corpora for their own benefit to look up information about language that is not available elsewhere. With the click of a mouse, corpus users can be empowered via the combined intuitions of hundreds of other language users. Although there are already a number of master's programmes in TESOL, TEFL or ELT that offer students optional modules in corpus linguistics, there do not seem to be many programmes where, instead of receiving training in general corpus linguistics, teachers are being specifically trained in applied uses of corpora for language teaching. This would be an important development if corpora are to become more relevant and present in everyday teaching.

Another need for the future is further development of corpora and concordancers for ELT. Some advances in this direction have already been made, as seen earlier in Cocetta (2011) and Widmann et al., (2011), although the corpora described in those studies are very small and have very limited uses. Another example is the *Compleat Lexical Tutor*, developed by Tom Cobb (Cobb, n.d.), which contains specific tools for creating data-driven learning exercises. For Better English (<http://forbetterenglish.com/>) developed by Kilgarriff et al. (2008), though originally conceived for lexicography, is a tool which automatically filters out concordances from a very large general English corpus so as to prioritise full sentences that are not too long or too short and also sentences exhibiting typical patterns of usage, while at the same time leaving out concordances which contain infrequent, more difficult words. SkELL (see Table 1), in turn, is a further development of the corpus-filtering technology developed in For Better English, presenting novice corpus users with an extremely simple and intuitive interface of the highly sophisticated Sketch Engine tool (Kilgarriff et al., 2004). In order for these and a number of other corpus tools and resources conceived for language teaching to be further developed, however, it is important that they should be tried out and tested by actual teachers and learners.

Beyond this, in the future it should also be possible to integrate corpora and concordancing software with other applications, such as CALL software and simple text editors. Indeed, some progress in this direction has already been made, such as the Concord Writer tool in the previously mentioned

Compleat Lexical Tutor, where learners can input their own text and link the words they write dynamically to concordances.

Conclusion

This chapter began by explaining corpora and corpus analysis tools. Examples of how different corpora can and have been used in ELT were given, and key points of debate were raised. The chapter concluded with some ideas for the future of corpora in ELT. While the growing influence of corpora seems to be undeniable, it remains to be seen whether one day they will become as essential to language teachers and learners as other, more conventional ELT materials and resources.

Discussion questions

- To what extent should language data from corpora be edited or simplified for pedagogic purposes?
- In what situations would it be appropriate to use parallel corpora in ELT?
- Discuss scenarios where the direct use of corpora by language learners can be useful in EFL writing.
- In your own professional context, to what extent is it realistic to expect teachers to develop corpus skills and use corpora with learners?

Related topics

ELT materials; Language curriculum design; Language teacher education; Learner autonomy; Language awareness; 'Native speakers', English and ELT; Questioning 'English-only' classrooms; World Englishes and English as a lingua franca.

Further reading

Flowerdew, L. (2011) *Corpora and language education*. London: Palgrave Macmillan. (A comprehensive overview of corpora in language education.)

Frankenberg-Garcia, A. (2012) 'Integrating corpora with everyday language teaching', in J. Thomas and A. Boulton (eds) *Input, process and product: Developments in teaching and language corpora*. Brno: Masaryk University Press. 36–53. (A discussion and examples of how the direct use of corpora can be integrated with everyday teaching.)

Gavioli, L. (2005) *Exploring corpora for ESP learning*. Amsterdam: John Benjamins. (A good introduction to corpora for those involved in teaching ESP.)

O'Keeffe, A., McCarthy, M. and Carter, R. (2007) *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press. (An account of using corpus data to produce ELT coursebooks.)

Online tutorials on the use of corpora in language teaching are available at

- http://calper.la.psu.edu/corpus_portal/tutorial_overview.php
- http://www.ict4lt.org/en/en_mod3-4.htm
- <https://eltadvantage.ed2go.com/>

References

Anthony, L. (2014) *AntConc 3.4.3*. Tokyo: Waseda University. Retrieved from <http://www.laurenceanthony.net/software/antconc/>

Baisa, V. and Suchomel, V. (2014) 'SkELL: Web interface for English language learning', in A. Horák and P. Rychlý (eds) *Proceedings of recent advances in Slavonic natural language processing*. Brno: Tribun EU. 63–70.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999) *Longman grammar of spoken and written English*. Harlow: Longman.

Boulton, A. (2010) 'Data-driven learning: Taking the computer out of the equation'. *Language Learning*, 60/3. 534–572.

Boulton, A. and Pérez-Paredes, P. (eds) (2014) *ReCALL special issue: Researching uses of corpora for language teaching and learning*. Cambridge: Cambridge University Press.

- Charles, M. (2012) 'Proper vocabulary and juicy collocations: EAP students evaluate do-it-yourself corpus- building'. *English for Specific Purposes*, 31/2. 93–102.
- Cobb, T (n.d.) *Compleat Lexical Tutor*. Retrieved from <http://www.lextutor.ca/>
- Collins COBUILD. (n.d.) *English for Learners*. Retrieved from <http://www.collinsdictionary.com/dictionary/english-cobuild-learners>
- Cook, G. (1998) 'The uses of reality: A reply to Ronald Carter'. *ELT Journal*, 51/1. 57–63.
- Firth, J. R. (1957) 'Modes of meaning', in *Papers in linguistics 1934–1951*. Oxford: Oxford University Press. 190–215.
- Frankenberg-Garcia, A. (2005) 'Pedagogical uses of monolingual and parallel concordances'. *ELT Journal*, 59/3. 189–198.
- Frankenberg-Garcia, A. (2007) 'Lost in parallel concordances', in W. Teubert (ed.) *Corpus Linguistics: Critical concepts in linguistics*. London: Routledge, IV. 176–190.
- Frankenberg-Garcia, A. (2012a) 'Integrating corpora with everyday language teaching', in J. Thomas and A. Boulton (eds) *Input, process and product: Developments in teaching and language corpora*. Brno: Masaryk University Press. 36–53.
- Frankenberg-Garcia, A. (2012b) 'Raising teacher's awareness of corpora'. *Language Teaching*, 45/4. 475–489.
- Frankenberg-Garcia, A. (2014) 'How language learners can benefit from corpora, or not' *Recherches en didactique des langues et des cultures: Les Cahiers de l'Acedle*, 1/1. 93–110.
- Frankenberg-Garcia, A. AND Santos, D. (2003) 'Introducing COMPARA: the Portuguese-English Parallel Corpus'. In F. Zanettin, S. Bernardini & D. Stewart (eds.) *Corpora in Translator Education*. Manchester: St. Jerome. 71–87.
- Gabrielatos, C. (2005) 'Corpora and language teaching: Just a fling or wedding bells?' *Teaching English as a Second Language Electronic Journal (TESL-EJ)*, 8/4. 1–35.
- Gavioli, L. (2005) *Exploring corpora for ESP learning*. Amsterdam: John Benjamins.
- Granger, S. (2003) 'The international corpus of learner English: A new resource for foreign language learning and teaching and second language acquisition research'. *TESOL Quarterly*, 37/3. 538–546.
- Johns, T. (1991) 'Should you be persuaded: Two samples of data-driven learning materials'. *English Language Research Journal*, 4. 1–16.
- Kennedy, C. and Miceli, C. (2001) 'An evaluation of intermediate students' approaches to corpus investigation'. *Language Learning & Technology*, 5/3. 77–90.
- Kilgarriff, A., Husák, M., Mcadam, K., Rundell, M. and Rychlý, P. (2008) 'GDEX: automatically finding good dictionary examples in a corpus', in E. Bernal and J. Decesaris (eds) *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra. 425–433.
- Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. (2004) 'The sketch engine', in *Proceedings of Euralex*. Lorient, France. 105–116.
- Lenko, A. (2014) 'Is this enough? A qualitative evaluation of the effectiveness of a teacher-training course on the use of corpora in language education'. *ReCALL*, 26/2. 260–278.
- Longman Dictionary of Contemporary English*. Retrieved from <http://www.ldoceonline.com/about.html>
- Macmillan English Dictionary*. Retrieved from <http://www.macmillandictionary.com/>
- McCarthy, M., McCarten, J. and Sandiford, H. (2005) *Touchstone. Student book 1*. Cambridge: Cambridge University Press. Retrieved from <http://www.cambridge.org/gb/cambridgeenglish/catalog/adult-courses/touchstone>
- McEnery, T., Xiao, R. and Tono, Y. (2006) *Corpus-based language studies*. London: Routledge.
- Mukherjee, J. (2004) 'Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany'. *Language and Computers*, 52/1. 239–250.
- Römer, U. (2004) 'Comparing real and ideal language learner input', in G. Aston, S. Bernardini and D. Stewart (eds) *Corpora and language learners*. Amsterdam: John Benjamins. 151–168.
- Römer, U. (2009) 'Corpus research and practice: What help do teachers need and what can we offer?' in K. Aijmer (ed.) *Corpora and language teaching*. Amsterdam: John Benjamins. 83–98.
- Scott, M. (2012) *WordSmith Tools 6.0*. Lexical Analysis Software and Oxford University Press. Retrieved from <http://www.lexically.net/wordsmith/>
- Seidlhofer, B. (2004) 'Research perspectives on teaching English as a lingua franca'. *Annual Review of Applied Linguistics*, 24. 209–239.
- Sinclair, J. (1991) *Corpus, concordance and collocation*. Oxford: Oxford University Press.
- Someya, Y. (1999) *A corpus-based study of lexical and grammatical features of written business English*. MA dissertation. University of Tokyo.
- Tognini-Bonelli, E. (2001) *Corpus linguistics at work*. Amsterdam: John Benjamins.
- Widdowson, H. G. (1998) 'Context, community and authentic language'. *TESOL Quarterly*, 32/4. 705–716.
- Widdowson, H. G. (2000) 'On the limitations of linguistics applied'. *Applied Linguistics*, 21/1. 3–25.
- Widmann, J., Kohn, K. and Ziai, R. (2011) 'The SACODEYL search tool. Exploiting corpora for language learning purposes', in A. Frankenberg-Garcia, L. Flowerdew and G. Aston (eds) *New trends in corpora and language learning*. London: Bloomsbury. 167–178. Retrieved from www.um.es/sacodeyl/
- Zipf, G. (1949) *Human behavior and the principle of least effort: An introduction to human ecology*. New York: Hafner.