Why I love corpora

In this article, corpus users tell us about the times when they thanked their lucky stars for corpora.

Susan Hunston ~ Paul Heacock ~ Anne O'Keeffe ~ Melissa Good ~ Jane Evison ~ Sheena Gardner ~ Ana Llinares ~ Regina Weinert ~ Tom Morton ~ Rachel Whittaker

Susan Hunston
Head of School of English, Drama and American and Canadian Studies, University of Birmingham, UK

I was 'converted' to using a corpus in the early 1990s, when I was working at the University of Surrey. I had invited Gill Francis to give a talk at the newly convened Linguistics research seminar. I had known Gill previously but did not know any details of what she was currently working on, other than it was in an organization called 'Cobuild'. Gill showed us work she was doing on the word 'possible', which is unusual because it behaves rather like 'probable', 'likely' etc ('It is possible that', 'a possible cause' and so on) but also like 'difficult' and 'easy' ('It is possible to' and so on) and also just like itself ('The best reason possible' and so on). The information itself struck me as only mildly fascinating (sorry Gill!), but the way it was presented, using concordance lines, took my breath away.

I sometimes say I fell in love with concordance lines at that moment, and it is only a little bit of an exaggeration. I adored the way the lines looked, the neatness of the patterning, and the way they made intuitive knowledge visible. I also loved the focus on individual words – it was as if words had a 'secret life' that was revealed using this technique. I knew immediately that whatever else I did with my life, I wanted to work with concordance lines. A

couple of years later I was lucky enough to land a job at Cobuild, and I've never looked back! Corpus techniques have moved on a lot since then, and scrutinizing concordance lines sometimes seems a little old-fashioned, but it is still my preferred way of working.

Paul Heacock

Publishing Manager, Dictionaries and Corpus, Cambridge University Press, US

My own 'thank goodness for corpus' moment came years ago when a colleague asked me if it was ok to say 'I don't like to eat alone' (rather than 'I don't like eating alone'). The question displays one of the key misconceptions about corpora – they never show you if something is right or wrong, just what it is most people do. It turned out that most people (most American English speakers, at any rate) prefer to follow 'I like' with a gerund (i.e. 'eating'), and not a 'to' infinitive. It's a somewhat unintuitive answer to me, which is exactly what makes corpora so valuable.

Anne O'Keeffe

Senior Lecturer, Mary Immaculate College, Limerick, Ireland

My story is about when I realized the power of the corpus in a funny sense. Years back, a senior colleague used to send out rather abrupt and direct emails. Around the same time, I was asked to give a staff seminar on my area of research. Because none of my colleagues knew what a corpus was, I wanted to give it all a context, so I decided to base part of it on a corpus of these emails. I called it the "mystery email corpus" and did wordlist analyses on it. I knew she wouldn't be there and it went down really well with the audience.

A few days later, she nabbed me and said that she'd heard about my talk. She wanted to know what I'd said but I just joked that it was anonymous data. I mentioned some of my findings about her use of pronouns. Her e-mails did actually become less abrupt as a result!

Melissa Good

Senior Development Editor for English Profile, Cambridge University Press, UK

My 'eureka moment' came when I was preparing for my first Cambridge University Press job interview and was learning about Cambridge dictionaries. I'd never heard of anything so cool! In my interview I was shown some concordance lines (for the word 'strength') and on my first day at work the Cambridge International Corpus in all its splendour was unveiled. So big, yet so easy to use.

Jane Evison

University Teacher, University of Nottingham, UK

Something changed when I made my first corpus as part of my Master's dissertation over a decade ago. I didn't even set out to make a corpus – in fact, when I put my dissertation proposal together, I don't think I even knew

that corpora existed. What I did know was that I wanted some evidence that successful users of English (Prodromou's SUEs – I didn't know that they existed then either) didn't regularly use clause-length openers such as 'that's a good point but' when they wanted to introduce an alternative opinion or viewpoint. I was teaching Cambridge FCE classes at the time, and a large proportion of my students seemed to be learning these gambits which seemed to fill up their working memories, leaving no room for them to formulate what it was they did think that was different from their partner, group or class. Once I'd got audio recordings and transcribed them, I discovered concordancing (I don't even remember how), and suddenly I was a corpus analyst. With concordances, I could very easily identify all the turn-openings and discover not only the almost total absence of the exponents taught in the books, but the endemic nature of the routinised 'yeah buts' which are the mainstay of our conversational 'toing and froing'.

Afterword: When I drafted this piece, I really wanted to use 'toing and froing' in the final sentence, but was concerned that, like so many spoken expressions, it looks a bit odd when it's written down. A quick search of the web (the biggest corpus of all) revealed 175,000 hits, and so it stayed. So, yes, life has never been the same since I accidentally made a corpus. And by the way, 'that's a good point but' had just 43,900 hits.

Sheena Gardner

Reader in Educational Linguistics, University of Birmingham, UK

A recent example involves the conjunctions 'ere' and 'whereat' which I read in a student essay, recognized, and thought - surely no-one uses that these days? I checked SketchEngine (see http://www.sketchengine.co.uk/) and was able to see where it appears on the Web (UKWAC), in the British National Corpus and in the British Academic Written English corpus. Most uses do seem to be archaic, and it turned out to be much less frequent than 'whereupon'.

I'm not sure I can remember my first use – I think Tim John's kibbizers (see http://www.eisu.bham.ac.uk/support/online/kibbitzers.shtml) certainly involved a eureka moment – in the way they help students identify collocations and evaluations in their writing.

Ana Llinares

Lecturer, Universidad Autónoma de Madrid

When I started thinking of a topic for my PhD, I knew I wanted to analyse the type of language that was used by young EFL learners in their classroom interactions. In order to obtain a complete picture, I needed to collect a spoken corpus of learner data in different types of immersion contexts, both with native and non-native teachers. The recent relevance of corpora and, in particular, learner corpora for studies on second language development encouraged us to collect our own corpora: the UAMLESC (Universidad Autónoma de Madrid Learner English Spoken Corpus) between 1998 and

2005, and the UAM-CLIL corpus, since 2006. Learner corpora have been mainly used for cross-sectional studies of formal features of learner language. The challenge that I foresee is for a wider use of longitudinal corpora, and also in combination with pragmatic and functional approaches to learner language.

Regina Weinert

Reader in Germanic Linguistics, University of Sheffield, UK

My first encounter with corpus work came 20 years ago when I took up a postdoctoral position in the Human Communication Research Centre at the University of Edinburgh. I had elicited data before for my PhD, but Jim Miller's project on the syntax of spoken English made the use of corpus data absolutely essential. We worked with the ready-made corpus of Scottish Spoken English, collected by Jim and Keith Brown. We were also involved in the transcription and analysis of the English Map Task corpus, the pivotal data which would be examined from all angles within the interdisciplinary centre – by phoneticians and other linguists, psychologists, cognitive and computer scientists.

The moment when I thanked my lucky stars for the existence of a corpus came when I wanted to look at German (which was part of my job). Since most of my days were taken up with the Map Task corpus, there was no time to collect let alone transcribe data. Fortunately, the Institut für deutsche Sprache (IdS) in Mannheim, Germany, already had two large spoken corpora, the DSK (*Dialogstrukturenkorpus*) and the FK (*Freiburger Korpus*). Nowadays, free access to the IdS data is only a few clicks away, and recently many transcriptions have been aligned with the sound; things weren't quite so simple in 1990. I remember writing letters to the IdS and applying for funding from the HCRC, which eventually led to me being able to print out reams of yard-wide paper on old printers with the precious data. I also remember big boxes of 'Tonbänder' arriving, reels a foot in diameter, which held the sound. It flummoxed the technicians, who could not resist the challenge of getting it all to work, but in the end they had to throw in the towel. I think my use of the data was limited to something like three years at the time and I have no idea if anyone ever rolled the reels back to Mannheim!

Tom Morton

Lecturer, Universidad Autónoma de Madrid

I first became aware of corpora when I was an EFL teacher. I thanked my lucky stars that here was tool which would let us have a glimpse of how English *really* worked. I might not have to make up all my examples when preparing for my classes! I got excited by the idea that I could get students interested in the adventure of exploring the language as it was used, and spent a good part of my lesson-planning time in the 90s thinking up activities starting from concordance output.

Later, when I got interested in researching classroom discourse, I began using a conversation analysis approach. My second 'eureka moment' came when it occurred to me (probably after a lot of other people!) that combining quantitative corpus work with Conversation Analysis could be a powerful methodology for understanding language use in educational contexts.

Rachel Whittaker
Lecturer, Universidad Autónoma de Madrid

Trying to solve a problem, we needed samples of texts of a certain type, by the group whose language we wanted to analyse. So the conditions for the creation of a corpus were there, but this was not the first objective. The aim was to learn about the problem we wanted to study. So the corpus was a by-product, but a logical one on the road.