Looking back or looking forward in corpus linguistics: What can the last 20 years suggest about the next?

Mike Scott
Aston University (United Kingdom)
mike@lexically.net

Abstract

Starting with a description of the software and hardware used for corpus linguistics in the late 1980s to early 1990s, this contribution discusses difficulties faced by the software designer when attempting to allow users to study text. Future human-machine interfaces may develop to be much more sophisticated, and certainly the aspects of text which can be studied will progress beyond plain text without images. Another area which will develop further is the study of patternings involving not just single words but word-relations across large stretches of text.

Keywords: history of corpus linguistics, MicroConcord, concordance, software design, collocational span.

Resumen

Mirando hacia atrás o mirando hacia delante en la lingüística del corpus: ¿Qué nos sugieren estos 20 años para las próximas décadas?

Esta aportación comienza con una descripción de los programas y los equipos informáticos que se emplearon en la lingüística del corpus a finales de los años 80 y comienzos de los 90 y seguidamente estudia las dificultades que han tenido que afrontar los diseñadores de programas informáticos para conseguir que los usuarios pudieran estudiar un texto. Cabe la posibilidad de que en el futuro se desarrollen interfaces hombre-máquina mucho más sofisticadas, y con toda seguridad se avanzará en aquellos aspectos textuales que puedan ser objeto de estudio superando el texto plano sin imágenes. Otro aspecto en el que se continuará avanzando será el estudio de los modelos o patrones que contienen no sólo palabras sueltas sino relaciones de palabras en tramos extensos de texto.

Palabras clave: historia de la lingüística del corpus, MicroConcord, concordancia, diseño de programas informáticos, colocaciones.

Introduction

The last twenty years have seen a revolution in the circulation of information and opinion just as important as the development of metal-smelting, of writing or of printing in previous epochs. An apparently simple change in technology (heating crushed rock, the creation of a character set, a machine to copy text quickly, a means of linking up millions of computers) brings about an enormous social change: it becomes possible to store knowledge so that generations can learn from their ancestors, to distribute it so that everyone in a community can know more, and now with the Internet, to enable discussion and exchange of ideas with very little regard for where or with whom one happens to be. Most of the psychological and social impacts of those technological changes could not have been imagined by the inventors and first users of those technologies. The first uses of writing were mostly for storing records concerning ownership and conquest but the effects soon began to include a spread of ideas and opinion and a sense of history; early printing centred on culturally-sanctioned official works but before very long led to challenges to the status quo, chiefly concerned with the right to freedom of speech. We are still in the very earliest years of the Internet revolution and we do not yet know all the changes it is bringing about.

One of them is a plague of zombies: public spaces full of people present in body but not in mind. About twenty years ago I had my first experience of Internet body/mind separation[1], sitting at my desk in my university office in Liverpool but mentally dislocated to an Australian university, reading documents stored on their servers. These online documents were highly factual, like the earliest Sumerian records from 5000 years ago (and about equally gripping). Terms like "online" and "server" were yet to be met and of course "social" did not yet collocate with "networking".

What follows is my own personal retrospective, leading I hope to anticipation of some possible future developments, but with the warning already implicit in what was written above: most of the interesting developments cannot easily be imagined in advance, even if they seem obvious in hindsight. That the motor car might lead to a network of surfaced

highways, with road signs and indeed with traffic accidents probably was predictable in the year 1896 or 1897, but Los Angeles' enormous suburban sprawl, the development of out-of-town shopping malls and the blight on town centres was not.

Changes

In the early 1990s most educated people had never heard of a "concordance"*;* the few that had associated it with study of religious text and an enormous amount of manual compilation. The word "concordancer" was even more restricted[2]. I had come across the form "concordancing" myself in the previous decade, thanks to Tim Johns: he and I were both officially concerned with English for Specific Purposes and that is really why we came into contact (Scott, 2012). However, Tim's enthusiasm for what micro-computers (as they were then known[3]) could be made to do matched my interest and so in the late 1980s we collaborated on a concordancer, which was published by Oxford University Press as *MicroConcord* (Scott & Johns, 1993): the element "micro" and the mid-word capital letter very much matching the spirit of the times.

Figure 1. *MicroConcord* (Scott & Johns, 1993), 5.25" floppy disk and F-key card.

Figure 1 shows 1993 technology: the large floppy disk which held the program, the manual, and a card to pace above the function keys on the keyboard. There were two floppies for different capacities of disk-drive; some users had no hard disk so had to run the program direct from the floppy.

At that time computer memory was very limited. *MicroConcord* required 200K of RAM – a standard micro-computer would have at least 640 but usually a whole megabyte.[4] Disk space was at a premium too: not only did some micro-computers have no hard drive at all, but costs were quite high. In October 1992, IBM launched the ThinkPad (mid-word capitals again) starting at US$4,350, with between 4 and 16 MB of RAM, a 10.4 inch colour display (640 by 480 pixels) and a huge 120MB hard drive (Stengel, 2012). That was a leading-edge computer, much more advanced than the university and lecturer machines for which *MicroConcord* was designed. In May 1992, Word Perfect 5.1 was issued; it cost US$495.[5] In 1994, Apple computer ran a magazine ad with a photo of what a bike messenger cum screenwriter would have on his (of course) Apple PowerBook (mid-word caps here too): "My first screenplay, a dictionary, a thesaurus, a spellchecker ... the number of a girl from my screenwriting class, a list of good bike repair shops, a detailed map of downtown, my résumé, my grocery list, Microsoft Word, the number of Ray Bazire who owes me money", etc. (Wichary, 2012). Clearly the PowerBook copywriters are suggesting that a computer does not only help users to do their work but also stores a variety of non-text resources and personal notes. 2012 advertising copy would be able to presume that many of these resources do not need to be stored on the user's computer but can be found online.

This little foray into a history which anyone aged 40 or more may remember, whether sharply or fuzzily, shows two things. First, the technology has changed very fast indeed, and second but more interestingly, user expectations and knowledge have changed importantly too.

In the early 1990s I was responsible, amongst other things, for introducing new overseas post-graduate students arriving at Liverpool University to the computer facilities they would have to use in their pre-sessional EAP classes. Students had to prepare a sizeable piece of academic writing concerned with their specialist subject, deliver a mini-presentation on the same topic, and design a poster for a poster session at the end of their 6-, 10- or 13-week course, so we wanted them all to feel confident with the University's computer systems. I do not think email was yet a requirement but the ability to type and print out one's academic writing was, because the student's own department would very soon require this too. At that time it was not uncommon for students from certain countries in the Middle East and South Asia not to know their way around the computer keyboard, so much hunting and pecking took place; students from the EEC (it became EU in 1993)

typically had some basic familiarity with the keyboard but still needed a lot of help in getting started with logging into the University computers, using email and getting started with word-processing.[6]

There were two main problems beside hunting and pecking. First, our University systems were somewhat complex and tricky, designed by computer engineers and not yet tuned to the knowledge and skills of ordinary users. Logging in (at the University of Liverpool in those days) assumed an awareness of the differences between Unix and Windows, since the university's core system was Unix but the software we ran was within a Windows environment running on top (not to mention another system to handle the network of different servers)! In the same way, starting a car engine in the 1900s required the user to get out a starting handle and crank it. The software of the 1990s was much less standardised than it is now, so users not only did not know what was possible, nor what each function is called, nor how to access it! Likewise, even as late as the 1980s you might meet someone who would claim to know how to drive a VW but not a Ford, because the basic functions were not yet standardised.

Second, many of my adult students were afraid to experiment, for fear of breaking something or looking foolish. The same problem I was very used to as an EAP and EFL teacher, risk-taking, but here for some amplified by the fear the humanities-trained student has of science and engineering.

At that time, the idea of a concordance was a tricky one. Any student (or worse because usually more conservative, any teacher) found it immediately daunting, because we are all trained to read linearly. To see a screenful of text where each line was unrelated to the line above and where words were incomplete at left and right edges of the KWIC concordance was very distracting and impeded face validity. Even now it is hard to learn to read a concordance vertically, sorting on the collocates at L1 or R1 position to gain an impression of word-patternings.

The problem of unfamiliarity has diminished greatly because we are all in the habit of using search-engines, and essentially these give a view which is rather similar to a concordance, with a set of unrelated entries, usually with our search-word highlighted and centred in the display.

Design problems

My own work in devising corpus linguistic software had to solve three kinds of problem:

- of making each function work as it should;
- allowing the user to know what the appropriate settings are and alter them; and
- explaining the point of each function.

Compare that with the designers of Microsoft Word, where most functions such as inserting a footnote or a picture with text flowing around it, or a section in italics, are ones which the literate reader and writer already knew about in printed essays, books, magazines, etc. even if they did not know how to carry them out in MS Word. And many of the technical terms "footnote", "italics" and so on, were already in standard use (admittedly users did have to learn a new meaning for the verbs "crash" and "back up"). Accordingly, MS Word's designers did not need to worry much about the third problem. I have found it in some ways the hardest of the three, in the sense that a lot of time needs to be spent on designing the Help system and then on software demonstrations, training workshops and follow-up email support. Empathy is required, an ability to guess what the other person imagines, knows and does not know; an ability all teachers need.

The first of the problems above is technical, it is not always easy but it requires patience and problem-solving, in thinking of ways to get something done despite limitations of machines. On the whole that has probably taken up about one-third of my time developing corpus software. In the early years the shortage of memory and disk space meant that ways had to be found to compute in the most economical way possible; this made it hard to read the code later and remember the various space-saving tricks employed – in other words, debugging and improving was made harder. In later years, software has become bloated because there is no longer the same need to find really economical ways of solving a problem. Moore's Law (Moore, 1965) has meant dramatic improvements in the speed of the chips inside the computer, the size of the disk drives and memory, which in turn has made corpus procedures work faster. We may now require more text to be processed and corpora have also grown in size, but still a procedure which ten years ago would have slowed things down unacceptably can now be allowed, so more

error-checking to avoid crashes can be brought in. At the same time, programmers like to find elegant solutions, and I will sometimes unnecessarily spend hours refining a routine in a search for elegance and efficiency simply because I want my code to satisfy me aesthetically.

The second problem concerns the human-machine interface and is much trickier. To make the way to operate a tool so intuitive that a user can succeed without looking for help is the aim,[7] and part of the solution is to keep things simple so that the burden of choosing does not overload the user. In the case of a pencil or a pen there isn't a problem anyway since the tool can only do one thing, but a multi-faceted tool like an automobile involves lots of choices about fuel, road traction, safety etc., so much so that we do not let people do it without quite a lot of training and basic skill testing. On the other hand, simplicity means strait-jacketing the user, and it seems silly to restrict what is possible merely because it is hard to show the user what the choices are.

Corpus linguistics in the Future

Let us now turn to the future. Will corpus linguistics grow as a discipline, or maybe even die out? What directions may corpus research follow?

Sound and video

Twenty years ago the PC already was capable of colour and sound, but corpus software typically only used three or four of 16 basic colours and no sounds except beeps indicating an error. Even now, corpora with sound and video accompaniment are very initial and sparse. There are many databases of isolated utterance recordings made for general computational linguistic projects (for example at the Linguistic Data Consortium or LDC[8]), but as these do not consist of normal running text or conversation, they do not contribute straightforwardly to corpus linguistics. The International Corpus of English (ICE) corpora do contain sound files (ICE-GB has a total of 70 hours of recorded speech, for example).[9] MICASE at the University of Michigan has 200 hours, just under 2 million words of Academic English.[10] The British National Corpus has 10m words of speech, but did not include the corresponding sound files.[11]

In the future, I believe ways could be found to transcribe corpora based on radio and TV programmes and movies. As always, copyright is the main

stumbling block, but many TV programmes are pre-scripted or transcribed in the production process so nearly all the work is done and technology to align the sound/video file with the text transcript is already available. Copyright is not necessarily a problem if the resulting video is sold, as the presence of DVD stands in supermarkets shows.

Individual corpora

One of the strengths of *WordSmith Tools* and similar software is the ability to process any corpus the user has access to, which can include official corpora like the BNC but, more interesting, corpora the user builds up him- or herself by downloading or institutionally from colleagues or students. As electronic resources on the web are increasing, it is very likely that more and more informal or home-made corpora will be built and used.

Corpus linguistics and other disciplines

It was already clear twenty years ago that corpus tools have a lot to offer language teachers (not that all teachers expressed interest, or in my opinion need to) but getting corpus resources used by students is still not a general practice in schools world-wide. With time, I foresee increasing expansion. However, the schools and colleges in most countries are still seriously restricted in computer (as opposed to mobile phone) availability and resources are simply not there for the purchase of software. It is likely that increasing use will be made of free software like *AntConc* (Anthony, 2012), which shares many facilities of *WordSmith*.[12] It is not likely that interest in language-learning will diminish over the foreseeable future, though the mix of languages studied will carry on changing. At the same time corpus resources and corpus tools will become standard resources. Not everyone will need to use them, for not every gardener uses a spade, but they will become standard tools used by a wider range of professionals: historians, biologists, medical and political science students, for example. That may mean that corpus linguistics itself ceases to be a discipline in its own right (there is no department of Spade Sciences that I'm aware of).

Certainly corpus tools will continue to develop. Corpus Linguistics is not just about data resources, such as corpora. It is about "adding value to data". That is, we might drown in a flood of data if we were not able to filter it and seek out patterns in it. For example, after early emphasis on single words or phrases of interest to a researcher, studying their Keyword-in-Context

(KWIC) contexts trying to establish or refute typicality, we have now moved further into much more focussed study of collocation. In other words, looking at how patternings within the context give a richer understanding of word patterning.

Collocation

Collocation must here be understood in its widest sense. The influential Osti Report study reporting on work carried out in the 1960s appeared to show that although "each node has an infinite region of influence, the influence decreasing the further away from the node you go …" (Sinclair, Jones & Daley, 2004: 48), it is very difficult in practice to find significant collocates of a word if one's horizons extend beyond four words to the left or right. Sinclair, Jones and Daley (2004) do not put it metaphorically, but it seems they viewed collocation as one might view magnetism, an attractive force which tails off rapidly with distance. On the other hand, armchair experimentation tells us that if say "eat" and "bananas" are collocates, it would be very possible for a large number of words or even clauses or sentences to be found between the two tokens, as in a story beginning "Let me tell you what to eat and what not to eat when travelling in …". Accordingly, perhaps we should think of the attraction between node and collocate as also sharing some qualities with gravity, which does not diminish with distance. Further, some space must be left for a kind of negative collocation, where a node shows a tendency to avoid a given collocate (much as human beings avoid each other). Hoey's theories (Hoey, 2005; but also in numerous other publications) stress that words keep or avoid company over much greater spans than just four or five words. Similarly we have seen increasing interest in multi-word units, n-grams, bundles, clusters, and concgrams (Cheng, Greaves & Warren, 2006).

Conclusion

Corpus Linguistics may or may not survive as a discipline but I am very confident that the ideas and resources built on foundations going right back to the 1930s will continue to develop and shape resources and tools. This will in turn keep corpus linguists and members of many other disciplines, and for that matter readers of *Ibérica*, busy for many years to come.

[Paper received 13 March 2012]
[Revised paper accepted 21 May 2012]

References

Anthony, L. (2012). *AntConc*. URL: http://www.antlab.sci.waseda.ac.jp/antconc_index.html [01/03/12]

Cheng, W., C. Greaves & M. Warren (2006). "From n-gram to skipgram to concgram". *International Journal of Corpus Linguistics* 11: 411-433.

Gibson, W. (1986). *Count Zero*. New York: Victor Gollancz.

Hoey, M. (2005). *Lexical Priming: A new Theory of Words and Language*. London: Routledge.

Moore, G.E. (1965). "Cramming more components onto integrated circuits". *Electronics* vol. 38 No. 8. URL: ftp://download.intel.com/ museum/Moores_Law/Articles-Press_Releases/Gordon_Moore_1965_ Article.pdf [01/03/12]

Scott, M. & T. Johns (1993). *MicroConcord*. Oxford: Oxford University Press.

Scott, M. (2012). URL: http://www.lexically.net/personal_pages/memories %20of%20Tim%20Johns.html [01/03/12]

Sinclair, J., S. Jones & R. Daley. [1970] (2004). *English Collocation Studies: The OSTI Report*. London & New York: Continuum.

Stengel, S. (2012). URL: http://oldcomputers.net/ [01/03/12]

Wichary, M. (2012). URL: http://www.aresluna.org/attached/computerhistory/ads/international/apple/pics/annual94-powerbook 5 [01/03/12]

Originally qualified as a language teacher, teaching English for the British Council in Brazil and Mexico, **Mike Scott** eventually moved to Liverpool University where he worked first in Applied Linguistics with emphasis on language teaching and English for Specific Purposes. A parallel interest in Corpus Linguistics and software design and development, however, eventually led to the publication of first *MicroConcord* and then *WordSmith Tools*. Nowadays he works and researches in Corpus Linguistics while maintaining the development of *WordSmith Tools* and supporting its extensive community of users in many parts of the world.

NOTES

[i] See the novels of William Gibson, for instance *Count Zero* published in 1986, for early awareness of this.

[2] Still is. A new copy of MS Word 2010 underlines the word in red.

[3] Computers at the time were either mainframe, mini- or micro-computers. The term PC had not yet come into common use, nor had its association with the Windows operating system. The size of the computer's box has now given way to whether it is placed on one's lap or one's desk, and though huge air-conditioned computers still exist these are likely to be called super-computers.

[4] A 2012 version of Windows with the 2012 *WordSmith Tools* would be uncomfortable with less than one thousand times as much RAM.

[5] See Polsson's *Chronology of Personal Computers* at URL: http://pctimeline.info/comp1992.htm

[6] Another very dated word: "word-processing" had come in with dedicated "word-processors" in the 1980s, that is specialised computers which could only be used for preparing documents, because computing still seemed redolent of white-coated technicians, air conditioning, expensive machinery.

[7] I am very aware that in this aim I get a very mediocre score!

[8] URL: http://www.ldc.upenn.edu/Catalog/topten.jsp

[9] For ICE, see URL: http://ice-corpora.net/ice/index.htm. For ICE-GB see http://www.ucl.ac.uk/english-usage/projects/ice-gb/index.htm

[10] URL: http://micase.elicorpora.info/

[11] They are being edited and may be available by the time you read this – see URL: http://www.phon.ox.ac.uk/SpokenBNC. But also see Dave Lee's *Devoted to Corpora* site for more sound archives at URL: http://tiny.cc/corpora

[12] And has been produced by Laurence Anthony in Japan with my blessing and encouragement since 2002.