In memory of Geoffrey Leech

It is with some trepidation that I approached the writing of this obituary for Geoff. The trepidation relates to the scale of the task — Geoff contributed so much to so many fields.

I first knew Geoff’s work in pragmatics and stylistics — like many others I only came across his corpus research after devouring *Principles of Pragmatics* and *A Linguistic Guide to English Poetry*. However, given the readership of this journal, I thought I would limit what I write here to Geoff the corpus linguist. This decision is aided greatly by the esteem in which his work is held in other fields — I have no doubt that in time obituaries covering his substantial contributions to pragmatics and stylistics will appear elsewhere. So, tempting as it is to discuss his work beyond corpus linguistics, here I focus on Geoff’s contributions to corpus linguistics and, in particular, his use of corpora to inform his work on English grammar.

I first really got to know Geoff in 1989. Our paths had crossed before, but it was at that time that his work with IBM on treebanking and automatic part-of-speech annotation gripped me to the extent that I started to talk to him about it. By happy coincidence, he also became interested in my PhD work on pragmatics at the time and, before long, I had moved back to Lancaster to work with him on a treebanking project for IBM TJ Watson Research Labs. In that work I also had the opportunity to meet and work with another founding figure in the use of corpora, Fred Jelinek. While at the time I took working with people like Fred and Geoff for granted, in hindsight I can only see in that attitude the casual folly of youth. I was immensely privileged to work with such people and to be able to interact with them. I was certainly enriched by it, as were all people who met and collaborated with them.

The key story that needs to be told in Geoff’s obituary, however, is not how I worked with him, nor how things were in 1989. The story that should be told is how the state of affairs in 1989 came about. In telling that story in this obituary I have decided to use Geoff’s own words. I was fortunate to interview Geoff in early 2014 on the question of the early history of his corpus work. I did this for a MOOC I am running for FutureLearn on corpus linguistics. Though retired, it was characteristic of Geoff that he generously agreed to be interviewed in this way. While the recordings at the time seemed routine, I now see the immense value of them — he
told a story that I did not fully know and which, in a very real sense, shaped him as a scholar and, through that, the field of corpus linguistics. Let us begin by addressing the question of when Geoff first came into contact with corpus data through his supervisor and colleague, Lord Quirk.

“Nobody, I think, before that time, before Quirk, had really thought of making a fairly representative corpus of the present-day English language” though this early corpus was “all paper-based”. Spoken material was gathered on reel to reel tape recorders worked on by “people like Jan Svartvik and Sidney Greenbaum, who were busy listening and transcribing this spoken language […] all this data was collected, and transcribed, and put into enormous metal cabinets which were like that because they had to be fireproof […] [the] corpus was there, in the room, in big metal cabinets. And so it wasn’t until the ‘70s, really, that that corpus at UCL was computerised.” By this time Geoff had moved to Lancaster, which was a very new university at the time, and started working on the LOB corpus — the genesis of which is somewhat surprising: “we sat around a table, about five of us who were specialising in modern English, and thought, what should we do? Lancaster is a new university, and can we do anything to put it on the map? And so a suggestion I made at that stage was, why not imitate this newfangled Brown Corpus”. Funding for the corpus was very limited to begin with: “we got a nice little grant from the publishers, Longman […] Because they had been in close liaison with Randolph Quirk […] so they were quite anxious to do the work. I think the amount of money was 3,200 pounds. A princely sum in those days”. Yet when work started it was daunting — and was very nearly abandoned: “The biggest task was actually the copyright problem we ran into. We were trying to negotiate copyright with the British publishers for about 500 texts. Eventually I sort of gave up the ghost. I said ‘I can’t do this anymore. We’ve run out of money. We’ve got rather primitive computing resources. We can’t get permission from the publishers, or the agents, or the authors. Let’s give it up.’ But at that stage, an angel from on high appeared in the form of Stig Johansson […] he was actually originally a student of Jan Svartvik’s in Lund, in Sweden. So he was already baptised into corpus studies. He happened to come to Lancaster on a Fulbright Fellowship or something for a year. […] he had the corpus bug, fortunately for us. He went back to Scandinavia and decided that he would take the Lancaster corpus, as it was then, to make this 1 million-word corpus which would be a matching corpus for the Brown Corpus in America. So we would then have two comparable corpora that we could compare and contrast […] Nelson Francis himself was actually instrumental in making us work towards that goal. Because when I said to him ‘we’re thinking of doing a corpus’, I sent him a letter. There were no emails or anything in those days. I sent him a letter saying ‘well, we’re thinking of building a corpus like your Brown Corpus in British English. Do you have any recommendations or thoughts about it?’ And so he said
‘well, for heaven’s sake, make it as close as possible a match for what we’ve done for American English. And of course that was a very sensible piece of advice. Because it did make […] possible [the] comparison of corpora across different regions and different generations.”

Having explained how the L and O of LOB came about, I asked Geoff about the B — Bergen: ‘when Stig Johansson got back to Norway, he was looking for somebody to collaborate with him. Got a quite generous grant from the Norwegian Scientific Foundation. But he needed somebody with computer expertise. And so he found it in someone called Knut Hofland in Bergen University. They had a kind of research centre there on language. And so he was able to tap into that and make a good collaboration between the two Norwegian universities. Which really made a big difference to the way the corpora developed”. Arising from these early links to Scandinavia came the ICAME organization, which has so enriched the field of corpus linguistics. Geoff Leech, Jan Svartvik and Stig Johansson were amongst the founders of the organization.

It was the need to undertake grammatical analysis on the corpus that led Geoff towards working with a team to create the CLAWS part-of-speech tagger. The team was: “Roger Garside from computing, me from the linguistics, and there were two research associates who we could afford to employ on that project. One of them was called Ian Marshall. Ian Marshall was playing around with the data, and he said ‘well, look, I’ve just discovered that if we do it this way, using this probabilistic algorithm, a sort of stochastic methodology, we can get a much higher success rate. Up, way up in the 90s, something like 96%’. And so that was a great breakthrough from our point of view […] after, of course, that became almost a kind of recognized methodology […] But there was also a kind of division between those who believed in the statistical methodology and those who believed in rule-driven methodology […] eventually I think people arrived at some sort of compromise. […] For some things it’s good to use rule-driven methods, for some purposes it’s good to use statistical methodology. The great trick is to, somehow, combine the two together”. It was through this work that Geoff started to work with the IBM team, a team in which Fred Jelinek was prominent: “He had this broad Czech accent, although he had been living in America for a generation. And he really thumbed his nose at the linguistic fraternity in those years. Chomsky? He wasn’t at all interested. He thought they were idiots, really. […] [Fred said]: ‘Give me enough data, and I can produce a much better machine translation algorithm than you can with as many rules as you’d like to build”.

This then was the context in which Geoff’s early contribution to corpus work began. Geoff was very clear as to why he focussed his corpus research on grammar rather than on other areas he was well known in. He focussed on “grammar because that was a suitable area to pick considering the corpora were rather small
[…] we could actually annotate grammatical features that we wanted to study”. Talking of the development of the Brown/LOB corpus to the present Brown family, Geoff acknowledged that “Over the years it’s been quite an international team, and all kinds of people have been involved”. Using such data “you’ve got the, say […] snapshots in principle for each language going through time. What type of changes in language do you think you can explore? […] it sounds a bit like stop motion photography”. Using such corpora, Geoff and his collaborators observed such features as “types of change which we thought were due to colloquialisation, as it’s called. That’s to say over the decades — I’m thinking later 20th century perhaps most of all — you’ll find that people are using more colloquial language. In other words, what is normal for speech is sort of invading the territory of written language […]. For example, the obvious example is the use of contracted forms like, I don’t think so instead of I do not think so […] they used to be considered taboo in written, serious written language. But now, of course, it appears in educational textbooks and even in some cases in academic writing and so forth. So that’s one change that has been very dramatic all the way through the 20th century up to now”.

Geoff’s story is a remarkable one. From the early UCL work through to the remarkable decision to build a corpus on the scale of LOB using very little money, he was at the very forefront of corpus development. That would have been enough to have ensured that his work would have been acknowledged in any review of corpus linguistics. Yet he then took that work and developed it, being present at the creation of both ICAME and viable part of speech taggers. He also reaped the benefits of his work building and annotating corpora through his grammars and writings on English grammar, in a series of truly remarkable publications. When one considers that all of this represented but one strand of the work in his career, the scale of Geoff’s impact on linguistics becomes apparent. I could have written this obituary about his work in pragmatics and not mentioned his work in stylistics or corpus linguistics. Similarly I could have written about his contribution to stylistics and ignored the other two areas. So I think it is fair to say that Geoff was one of the greats of linguistics. He will be sorely missed by all who knew him.

Tony McEnery, Lancaster University, England