The Development of ICAME and the Brown Family of Corpora

Geoffrey Leech[1] *

[1] Lancaster University

Abstract

Knut Hofland has been closely and continuously connected with two seminal developments in the history of corpus linguistics: the development of the organization known as ICAME (in full: the International Computer Archive of Modern and Medieval English), and the development of the Brown Family of Corpora. The best way I can find to pay tribute to Knut's key contribution to corpus linguistics is to sketch the history of these two interconnected development of corpus linguistics, and of his role in them.

Keywords: ICAME; Brown Family

*** Principal contact:**
Geoffrey Leech, Emiritus Professor,
Department of Linguistics and English Language, County South, Lancaster University, Lancaster, United Kingdom
Tel.: +44 1524 593036
E-mail: gleech@lancaster.ac.uk

1. The Early History of ICAME

It can truly be said that the genesis of ICAME was due to the creation of two corpora which eventually evolved into the 'Brown Family'. When I moved from University College London to the new University of Lancaster in 1969, the first research initiative of any importance I became involved in was the building of a 'Lancaster Corpus', which was planned to be a British equivalent of the Brown Corpus, created at Brown University in the USA in the early 1960s. In 1970, with the help of colleagues, I set up a research unit with the title CAMET (Computer Archive of Modern English Texts). Our goal was to make a collection, rather grandly called an 'archive', of corpora of which the initial members were to be the Brown Corpus, the newly launched Lancaster Corpus, and a computerized version of the Survey of English Usage corpus at UCL. We obtained some modest funding to enable us to begin the 'Lancaster Corpus', but we encountered increasing problems, of which the most troublesome were primitive computing facilities, our lack of computational expertise, and above all, difficulties of copyright clearance.

Stig Johansson was our saviour. As what we would now call a 'post-doc', he had come to Lancaster on a Leverhulme fellowship from Lund around 1975, and despite all our problems he was bitten by the corpus bug: when he returned to Scandinavia and was appointed to a docentship – then a chair – at Oslo, he obtained Norwegian funding that enabled him to complete the Lancaster Corpus – now appropriately renamed the Lancaster-Oslo/Bergen[1] Corpus, from the three universities jointly involved in its creation. The LOB Corpus – to give it its abbreviated name – was completed in 1978, and the first two corpora of the set of corpora subsequently known as the 'Brown Family' were in business.

The engagement of the University of Bergen in this enterprise was crucial: at that time, Stig, like myself, had little experience of computers, but was happily able to enlist the help of the *NAVFs EDB-senter for humanistisk forskning* – otherwise more simply known by its English title the *Norwegian Computing Centre for the Humanities* (*NCCH*) - to provide computer resources and know-how; and that Centre was located at Bergen. The meeting to found ICAME – then called the International Computer Archive of Modern English – was held in Oslo in 1977, and the 'founding fathers' (Nelson Francis, Stig Johansson, Arthur O. Sandved,[2] Jan Svartvik and myself) were joined at this first meeting by Jostein Hauge, Director of the NCCH, who was supportive in lending us the expertise and the facilities of the Centre to set our corpus plans in motion. We owed a lot to Jostein Hauge as head of the Centre, but it was Knut Hofland who was the computer engineer and programmer whose enthusiasm and versatile expertise really enabled our plans to come to fruition. From his youthful appearance, I assumed that Knut had only been working for the University for a short time – but his presence 'in the back room' made all the difference to the success of this, probably the world's first modern electronic corpus-building organization.

The intimate connection between the LOB Corpus and the founding of ICAME came about as follows.[3] Around 1976, the corpus project at Lancaster had run into the mire: when we attempted to obtain free permission to use the 500 two-thousand-word extracts from British English texts published in 1961, which we needed if our corpus was to match the Brown Corpus, the British publishers were loath to agree to this, and many demanded fees that we could not afford. Eventually we gave up the task of persuading them, as it seemed clear that the publishers

[1] The spelling *Lancaster-Oslo/Bergen* was due to Stig, and many, like me, must subsequently have wondered 'Why the hyphen between *Lancaster* and *Oslo*, and the slash between *Oslo* and *Bergen*?' It was only in 2009 that Stig revealed to me his thinking: the spelling, I believe, was intended to signify the embedding or one collaboration within another. That is, the main collaboration was the Anglo-Norwegian one, and within the Norwegian one there was another collaboration between Oslo and Bergen. This can be made clearer by bracketing as follows: *[Lancaster]-[[Oslo]-[Bergen]]*.

[2] Then Professor of English Language at the University of Oslo.

[3] This narrative is to be found in more detail, with documentation, in Leech and Johansson (2009).

were in touch with one another, and had decided collectively that they were not going to grant this free benefit to an obscure northern English university that few of them had heard of. (Lancaster University at that time was only about ten years old.) Stig, when he took the incomplete corpus back to Norway with him, could write to these publishers in the persona of the 'secretary general' of an important-sounding international organization wishing to include their texts in an archive of the English language for future global research. This made the granting of permission seem an honour to the copyright holder, rather than an imposition! The '*I*' in *ICAME* was therefore crucial in giving the archive a gloss of world importance. This was the immediate motivation for setting up ICAME – and, despite difficulties, the strategy worked, and Stig and Knut were able to finish the British counterpart of the Brown Corpus. ICAME was founded in February 1977, and in 1978 the LOB Corpus was completed.

Actually, Stig's title, as the main academic organizer of ICAME, was not 'secretary general', but the humbler title of 'co-ordinating secretary'. ICAME was a new style of democratic minimalist organization. By this I mean that it had no constitution, no membership, no president or chairman, no subscription, no administrative committee or executive board. For many years Stig and Knut between them ran the organization, with some volunteer help from keen members of the then tiny community of corpus linguists. Yet Stig was modestly styled as a co-ordinator. ICAME was an early example of the way the electronic revolution brought about a new kind of research community: a community bonded through electronic means – by email and the web – rather than through the traditional academic paraphernalia of constitutions, governing committees, subscriptions, presidents, treasurers, and the like.

Yet in the next two decades, ICAME evolved into a fully-fledged research community, an active force in the astonishing transformation of corpus linguistics over recent decades from a derided fringe group to an academic mainstream. In those early days, the activities of ICAME were threefold, all three enacted or overseen by Knut:

(a) The distribution of copies of the corpora (on magnetic tape) and of spin-offs from the corpora, such as concordances, which were distributed on microfiche.
(b) The organization (from 1979 onwards) of an annual conference.[4] The host of each conference informally took on the task as decided at the previous year's conference, and it was assumed that each conference would be financially self-supporting.
(c) The publication of a newsletter (*ICAME News*), which eventually, in 1987, morphed into *ICAME Journal* – although the numbering continued from *ICAME News*, the first issue of *ICAME Journal* being No. 11.

The change from *ICAME News* to *ICAME Journal* was a significant one, signalling that the members of the small but tightly-knit ICAME community were no longer content with a newsletter: they were already publishing corpus linguistic articles of academic significance, and they needed these to be published in a proper academic journal, alongside established academic journals on English and other modern languages, such as *English Studies* or *Anglia*. Since then *ICAME Journal* has appeared every year, and every issue was produced by Knut up to relatively recently, when production passed to Leeds and then to Lancaster. This change of title also signalled that ICAME, despite its name, was ceasing to be simply an archive (and a distribution centre) for corpora. It was becoming an academic community with its own conference and journal – and members were presenting and publishing papers not just on corpora, but on the

[4] In one year (1980) no conference was held. For the list of ICAME conferences, their locations and their published proceedings, see David Lee's corpus website (http://tiny.cc/corpora). A copy of the list can also be downloaded from the ICAME website.

research findings coming from corpus-based research. *ICAME Journal* was demonstrating the value of electronic corpora as the basis for research.

Another indicator of the 'democratic minimalism' of ICAME was that up to about 1987 Knut distributed a free copy of the *Journal* to anyone on the ICAME mailing list, or anyone who requested a copy. But financial cut-backs eventually forced him to charge a subscription for the journal – a subscription that was later added on to the conference fee, so that all attendees at ICAME conferences received a copy in their conference pack. Latterly the journal has been published on the Web, as well as on paper, and past numbers can be read or downloaded without payment.

In the 1990s came a reform: ICAME got a constitution, and an executive board with regulated membership. Knut became a permanent 'technical secretary'. But there was still no general membership or subscription.

2. ICAME and the publication of corpus linguistic material

Apart from the newsletter, in the early days Knut and Stig entered into the business of printing and publishing books: in 1982, two books were printed and published by the Centre in Bergen: Hofland and Johansson, *Word Frequencies in British and American English* (later taken over by the British publisher Longman) – the first of a number of printed word frequency dictionaries derived from the corpora; and Johansson (ed.) *Computer Corpora in English Language Research* – a set of research reports based on the second ICAME conference at Bergen. In publication terms, these were the first fruits of the ICAME enterprise. Bergen (i.e. Knut) continued to produce users' manuals for the corpora, but it soon became clear that the impact of research would be more advantageously spread through established academic publishers. So when, in 1989, Stig and Knut published their frequency lists based on the POS-tagged Brown and LOB Corpora, it was under the prestigious imprint of the Clarendon Press at Oxford. Also, after 1982, proceedings of ICAME conferences gradually became regular book-length publications, mostly published by the Dutch publisher Rodopi. The first Rodopi volume, edited by Jan Aarts and Willem Meijs, appeared in 1984, after the fourth ICAME conference in Nijmegen – the first to be held outside Scandinavia. It was entitled *Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research*, and incidentally its main title was a first step in popularizing the term 'corpus linguistics'.

3. ICAME conferences

The first two conferences of ICAME were held in Bergen, but in later years, the annual ICAME conference moved from one country to another, as the ICAME community gradually expanded. The heartland of ICAME remained Europe, and within Europe, North West Europe, or more specifically Scandinavia. But further afield, conferences have been held in Australia, Canada, and the USA. Bergen still holds the record for the largest number of conferences in one location – three. The UK holds the record for the country with the largest number of conferences – eight. And Knut probably holds the record for attending the largest number of ICAME conferences: he has attended, I believe, all except in one year when illness in the family kept him at home. In fact, Knut has been (in the best possible sense) the 'archorman' of ICAME. Although the centre at Bergen where Knut works has changed its name and its functions rather frequently (NCCH – HIT Centre – Aksis Centre – Uni Digital), Knut has remained the reassuringly constant element. Especially since the much-lamented death of Stig Johansson, he has provided the sense of continuity, linking the present ICAME with its past, and indeed its very beginning.

Unsurprisingly, since the 1990s, ICAME has gone online, and through Knut's expertise, its services to the corpus-using community have increasingly been available on the internet. It is worth now turning to the various ICAME services and resources that can be accessed through

the website (http://icame.uib.no/). In doing this, we will retrace some of the milestones of ICAME's development, and Knut's role in them.

4. ICAME corpora

A key part of the original plan was that ICAME would be a distribution centre for English language corpora. At the beginning a nucleus was provided by Brown, LOB and the London-Lund Corpus (Jan Svartvik's computerized version of the spoken data of the Survey of English Usage), but as more corpora came on board, they were added to the list of those available from Bergen. The first version of the ICAME Corpus Collection was distributed on CD from 1991. The second version, which can be ordered from the website, contains 20 corpora and their user manuals. However, it must be admitted that in the present technological climate, where corpora can be accessed online, or can be easily downloaded with or without licence, other sites such as the Oxford Text Archive and the Linguistic Data Consortium (LDC) have taken over the major part of this function of corpus archiving and distribution which had seemed most important at the founding of ICAME.

Of the corpora most closely associated with ICAME, mention must be made not only of Brown and LOB, but of other corpora compiled according to the same modal: the Australian Corpus of English, and Wellington Corpus of Written New Zealand English, and the Kolhapur Corpus of Indian English. But even more significant, from my point of view, were the two corpora created in Freiburg by Christian Mair and Marianne Hundt: the Freiburg LOB Corpus and the Freiburg Brown Corpus (affectionately known as FLOB and Frown – see 'the Brown Family of Corpora' below).

4.1 The ICAME Bibliography

It will seem unbelievable to many today that in the earlier days, it was possible to list virtually all the publications making reference to English language corpora. Later, Bengt Altenberg of Lund University undertook the increasingly onerous task of keeping track of new publications. The bibliography began as a page or two in Stig's 1982 volume, but by 1991, an updated version of Bengt's ICAME Bibliography was available on the ICAME file server at Bergen, and was also included in a book edited by Stig and Anna-Brita Stenström (*English Computer Corpora: Selected Papers and a Research Guide*)[5] published by Mouton de Gruyter. The bibliography at that time listed a set of five corpora that were at the basis of most published research, and which were flagged for individual publications: BCE, BUC, LOB, LLC and SEU (the Birmingham Collection of English Text, the Brown University Corpus, the Lancaster-Oslo/Bergen Corpus, the London-Lund Corpus and the Survey of English Usage). Eventually it was clear that any attempt at comprehensively listing English language corpus publications could not be maintained, and Bengt's bibliography became a self-updatable bibliography on the ICAME website, where on the 'wiki' principle authors can themselves add items to it.

4.2 The Corpora discussion list

Since 1995 **a**nother valuable service provided by Knut's ICAME Website has been in existence: the Corpora List, circulated by email, to which anyone interested in corpora can freely subscribe and contribute. In the intervening years this has become massively successful in attracting readers and contributors. It is a discussion platform on which information (for example, about new corpora, software, jobs and conferences) can be advertised, and all matters of interest to corpus linguists can be debated. An archive of posts and threads since 1995 can be consulted on the website. The Corpora List has now spread its influence well beyond it original remit. It not only deals with corpora for any language, but the topics it covers include computational linguistics, natural language processing, electronic dictionaries and the like. The diversification

[5] The 'Selected Papers' in this volume were presented at the 1989 ICAME conference, held at Bergen. The 'Research Guide' included not only Bengt's bibliography but a survey of English language corpora (only 36 were listed) and a survey of concordance programs by Knut Hofland.

of its readership is one clear testimony to the way corpus linguistics has spread its influence into all aspects of linguistics, computing and their interface.

4.3 The Brown family of corpora

After stressing the intimate connection between the founding of ICAME and the Brown and LOB Corpora, it is worth devoting a paragraph to the way Brown and LOB became the foundation of a whole 'family' of corpora of American and British English. But first, a word must be said about POS-tagging (or grammatical tagging, as it was generally called then). The Brown team under Nelson Francis and Henry Kučera achieved a world first when they completed the POS-tagging of the Brown Corpus, using a program (TAGGIT)[6] which correctly tagged 77 per cent of the words, the remainder being manually disambiguated. When I attended the first ICAME conference at Bergen, in a drinking session on the Bryggen, Nelson and Henry somehow found themselves agreeing to allow us to use their tagged Brown Corpus as a training corpus for our own probabilistic LOB Corpus tagger, which eventually became Roger Garside's tagging software, known as CLAWS (Constituent-Likelihood Word-tagging System). The term 'training corpus' was not current then, but being allowed to use the tagged Brown corpus to provide probability estimates for our own tagger, we (the Lancaster team) stumbled across the importance of this notion: the success of CLAWS was 96-97 percent, which was well in excess of the success of TAGGIT, although the remaining 3-4 per cent of the tags were erroneous and had to be corrected by hand. I should make it clear, though, that CLAWS was created as part of a joint project, and the work was partially completed at Oslo and Bergen.

The notion of building comparable corpora – that is, corpora which match one another as precisely as possible except for one key variable – began to take off when Nelson Francis advised me, in a letter of 1969, to build the projected Lancaster Corpus following exactly the design and sampling practice of the Brown Corpus. The corpora were selected from texts of the same date (1961), thereby making exact synchronic corpus comparisons of written American and British English possible. In the early 1990s, these two corpora became again the model for two additional comparable corpora, the Freiburg-Brown ('Frown') and the Freiburg-LOB ('FLOB') corpora developed by Christian Mair and Marianne Hundt, with texts sampled for the years 1991 and 1992. Later Nick Smith and I collaborated with Christian and Marianne in the POS-tagging of these two newer corpora, and in using comparable corpora, this time *diachronically*, to show what grammatical frequency changes had taken place between 1961 and 1991/2. Further developments took place with the completion of comparable corpora for earlier and later periods (1931, 2006).[7] The generation-gap of thirty years between one comparable corpus and another (1931, 1961, 1991/2) naturally suggests the analogy of a family with siblings, parents and grandparents: hence the name 'the Brown Family of Corpora' has become familiar, referring to seven corpora covering the period between 1931 and 2006.

5. Concluding remarks

I have focused almost complete on English language corpora in this retrospective survey. It is reasonable to claim that most of the pioneering work leading to the development of electronic corpora and of corpus linguistics started with English. But of course, English never had a monopoly of corpus linguistics, and the field has since widened to include a large number of the world's languages. ICAME has broadened its scope to include other languages as well as

[6] The TAGGIT program is described in Greene and Rubin (1971). For a more accessible account, see Garside et al. (1987: 32-3; 42-5).

[7] The set of corpora known as the Brown Family consist of the following: BLOB-1931, LOB, FLOB and BE06 for British English, and Brown, Frown and AmE06 for American English. There is only roughly a 15-year gap between the latest corpora (BE06 and AmE06) and FLOB and Frown. On the Brown family, see Smith and Leech (forthcoming 2013); and the BE06 corpus, see Baker (2009). Two at present incomplete corpora (BLOB-1901 and BBrown) are expected to join the Brown family when complete.

English:[8] and much research undertaken by Stig with Knut's help in later years involved the development and investigation of multilingual corpora, particularly the English-Norwegian Parallel Corpus. Also, diachronically, from the period when Matti Rissanen and his Helsinki team began to develop historical corpora in the mid 1980s, ICAME has extended its range in time, and the 'M' in its acronym now stands for both 'Modern' and 'Medieval'. ICAME has thus kept abreast of new developments, spreading well beyond its original bounds.

It has been fortunate indeed for the development of corpus linguistics over the past thirty-five years that Bergen became a headquarters for international corpus research under the banner of ICAME, and that Knut began his career as a corpus technologist there in the 1970s, continuing that same career. ICAME, the oldest organization for corpus linguistics, could not have developed as it has without the unique expertise and personal qualities of Knut Hofland.

References

Aarts, Jan and Willem Meijs (eds.) (1984) *Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research*. Amsterdam: Rodopi.

Baker, Paul (2009) *The BE06 Corpus of British English and recent language change.* International Journal of Corpus Linguistics, 14 (3). pp. 312-337.

Berndt, T. J. (2002). Friendship quality and social development. *Current Directions in Psychological Science, 11*, 7-10.

Garside, Roger, Geoffrey Leech and Geoffrey Sampson (eds.) (1987) *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Hofland, Knut and Stig Johansson, Stig (1982) *Word Frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanities; London: Longman.

Johansson, Stig (ed.) (1982) *Computer Corpora in English Language Research*. Bergen: Norwegian Computing Centre for the Humanities.

Johansson, Stig and Knut Hofland (1989) *Frequency Analysis of English Vocabulary and Grammar*. 2 vols. Oxford: Clarendon Press.

Johansson, Stig and Anna-Brita Stenström (eds.) (1991) *English Computer Corpora: Selected Papers and a Research Guide*. Berlin: Mouton de Gruyter.

Leech, Geoffrey and Stig Johansson (2009) 'The coming of ICAME', *ICAME Journal*, 33, 5-20.

Smith, Nicholas and Geoffrey Leech (forthcoming 2013) 'Verb structures in twentieth-century British English.' In Bas Aarts, Joanne Close, Geoffrey Leech and Sean Wallis (eds.) *The Verb Phrase in English: Investigating Recent Language Change with Corpora.* Cambridge: Cambridge University Press, pp. 68-98.

[8] The ICAME constitution states that one of its purposes is 'to include in its remit corpus-based studies of other languages, where English is a major comparative element'.