The Helsinki Corpus Festival (Helsinki, 27 Sept. – 2 Oct. 2011)

Conference Report – by Gabriella Mazzon

It was only too appropriate that the 20th anniversary of the Helsinki Corpus should be celebrated in style; let's admit it: this enterprise was the starting point of a revolution in most lines of study within our field. With the dramatic acceleration in corpus production and exploitation in the last few years, we have nearly forgotten what our papers looked like before digitalised collections and searching tools changed our perspectives.

It was with some trepidation that I went back to Helsinki, so many years after the memorable ICEHL in 1990, where we first realised the potential that the "Helsinki School" was starting to express. The city was as appealing as ever, and the "School" is also as lively as ever, the founding members seeming to have hardly aged, and the new members as eager and enthusiastic as the older ones still are.

Certainly, times have changed, and this event showed that corpus studies are continuously being re-discussed, questioned, refined and developed. This was testified not only by the numerous methodologically oriented contributions, but also by the posters and displays announcing new corpora or new encoding procedures being tested. A space was provided to announce the new XML version of the Helsinki Corpus, re-encoded according to the TEI guidelines; this testifies to the fact that the Festival was not only the celebration of an achievement, but also the announcement of innovation being planned and produced. The main common point seemed to be the need to integrate corpora or sections of various corpora to build up tailor-made databases; this requires a more standardised encoding system, in order for corpora to be able to "talk to each other".

The Festival proper was preceded by a workshop on historical pragmatics, one of the most vital areas at present, hosted by Irma Taavitsainen and Carla Suhr, who also kindly offered a warm-up buffet supper on the previous evening at the historic VARIENG premises, where we could be introduced to the very place "where it all started", while at the same time passing offices where younger co-workers were still busy on last-minute additions to the Conference materials or on their presentations. This workshop was indeed in such high demand that it had to be prolonged to the next day, thus overlapping with the event proper. It included a plenary talk by Andreas Jucker, who investigated the possibility to reconstruct meta-information on the perception of politeness and polite verbal behaviour through the Helsinki Corpus, trying to map the emergence and paths of words like *courtesy* through the Corpus' subsections, very much like fishermen try to trace the movements and the density of their "catch".

At the official Festival opening, after some initial reminiscent words by the "father" of the Helsinki corpus, Matti Rissanen, we were treated to a beautiful spell of choral singing that put everyone in an even more positive mood. Off we scattered to the various sections, already feeling regretful for what we had to miss: the programme was literally packed with exciting novelties, even the more minute case-studies an opportunity to illustrate new corpora, new applications, new tools. The people who structured the first corpora, and those who have used

them over the past twenty years, are now eager to do new things with them, to build new ones and to have them interact with each other – the historical corpus study agenda started to fall into shape. The main drive, on the one hand, is the need for standardisation, for creating bases for comparability but often from a modern perspective, the attempt to make units measurable and comparable. On the other hand, however, there is the urge of going back to the text, the idea that the distillation of data from corpora might not be all there is to the understanding of linguistic analysis. In the same way as some decades ago it was necessary to urge the scientific community to avoid excessive reliance on editions, through highlighting their limitations and faults even at the cost of admitting mistakes, and to go back to manuscripts as the "real" source of evidence, similarly now there is a tendency to grow suspicious of the pitfalls of "filtered" evidence. This can take the form of corpus mixing, dismembering, meta-analysing, pre-digesting and variously cross-fertilising. Annotation techniques got quite a share of attention, as well as new software being developed not only at Helsinki itself, but in various universities, which presented their projects also in a poster and software-demonstration session.

The second plenary featured Geoffrey Leech, another "founding father", who went on to exemplify the potential of tracing disappearance, as opposed to emergence – the obsolescence of a feature as opposed to its spread – something that, according to him, we do not do often enough, but that can also be of help. This was demonstrated through the relatively short-term diachrony of the growing "Brown Family" of corpora.

In the third plenary, Dawn Archer discussed various aspects of corpus annotation and searching tools; this reminded us that corpus construction is only the first step, and that the way in which the corpus is then analysed and pre-processed can significantly influence the results of the study, since choices are made that can also have a bearing on ensuing use of the corpus for different aims; the advantages, however, override these dangers, according to Archer and to most other participants. "Corpus annotation" was indeed the theme of one of the day's parallel sessions, focussing on spelling, genre, syntax, information structure and other types of meta-information being added to corpora – the closing discussion, in which the present writer played the devil's advocate, testified to the progress being made in this area, and also to the interest in the use of annotated corpora as heuristic, but also as teaching, tools. Unfortunately, an interesting workshop on "Metadata and Descriptors in Historical Texts" was run parallel to this session, in which the potential for integrating and coordinating different corpora was discussed. The third theme session of the day investigated the relationship between "Medieval English and Latin" through corpora.

The last plenary was given by Merja Kytö, another key figure in the compilation of the Helsinki Corpus, and more recently in dialogue studies. She introduced the element of geographical variation in historical corpora, giving a rich overview of the availability of materials for various areas, and critically examining the resources for a possible, and indeed much needed, "history of varieties of English", with particular attention to the problem of the authenticity of data. Again, a whole ensuing session was devoted to "Areal and regional variation in English historical corpora", with projects being illustrated starting from ME and ranging from this to Modern Irish English and to the on-going construction of the ICE - Singapore component. The parallel sessions, were, in this case, not thematic, and concentrated

more on case studies of individual words or word-classes, and of text-types respectively (the latter also featuring Daniela Cesiri on “Historical ESP”); furthermore, there was a workshop on the eighteenth-century grammatical tradition database. At the end of this day, the sumptuous conference dinner was a great opportunity to exchange more ideas and conversations, and to catch up with old friends.

The Sunday morning was also packed with interesting contributions, which unfortunately the present writer was unable to attend. Nevertheless, the impression was certainly that of a dense, exciting conference, bubbling with new projects and ideas – certainly much more than a celebration of past achievements, an acknowledgement of the starting point (and the starting people) of English historical corpus studies; it was also the announcement of a whole galaxy of new projects, involving new people, all testifying the vitality of this field.