HIT and ICAME
– A visiting researcher's observation

Wang, Jianxin
Beijing University of Post & Telecom, Beijing, China

Abstract. This short paper discusses very briefly some of the corpus-and-database-related research at the Humanities Information Technologies (HIT) Centre and HIT's relation to the International Computer Archive of Modern and Medieval English (ICAME). It also aims to clarify and update the relationships between ICAME and the ICAME COLLECTION, ICAME and HIT, and HIT and the Norwegian Computing Centre for the Humanities (NCCH).

The Humanities Information Technologies (HIT) Centre**,** established at the University of Bergen (UoB) on January 1, 1998, is a research centre which explores, develops and applies information technologies to different fields of the humanities. It was created through the convergence of three previously strong and independent units: the Norwegian Computing Centre for the Humanities (NCCH), the Norwegian Term Bank, and the Wittgenstein Archives. Before the HIT Centre was created, NCCH was an independent research institute set up in the early 1970s. Since the HIT Centre was founded, NCCH has become a part of HIT, and thus a part of the UoB. NCCH no longer exists independently. The Norwegian Term Bank and the Wittgenstein Archives are two large research projects; the integration enabled the HIT Centre to use their resources more efficiently.

In the field of corpus linguistics, NCCH was closely linked with the Lancaster-Oslo/Bergen Corpus (LOB), one of the two pioneering English language corpora in the early development of computer corpora. NCCH was responsible for editing part of the LOB corpus and tagging about half of the texts, using the CLAWS tagger developed at Lancaster University. Its contribution to the LOB corpus and the ensuing corpus-based research publications established the historically significant position of the NCCH centre in the computer corpus development.

NCCH has been an active member and supporter of the International Computer Archive of Modern and Medieval English (ICAME), which is an unofficial and informal international organization of linguists and information scientists. Initiated in 1977 by several scholars from England, the United States, Sweden, and Norway, ICAME has grown into one of the most influential organizations in English corpus linguistics, with a distinguished advisory board of 17 members and more than 100 participants at some of its annual conferences. It organizes the activities as listed below.

a. Collecting and distributing computerized English language corpora in the form of the ICAME COLLECTION.
b. Holding an annual conference since 1979 in different countries on English corpus linguistics . This year it will be held in Sydney, Australia.
c. Compiling and publishing a yearly *ICAME Journal* up until the present issue.
d. Maintaining an archive of corpora for research and purchase.

Among the above four activities, at least three – a), c), and d) – have been carried out at the NCCH, now HIT Centre, ever since the founding of the ICAME organization. It is probably not an exaggeration to say NCCH has in effect played the role of the standing centre for ICAME.

The ICAME COLLECTION of English Language Corpora (ICAME COLLECTION) is a collection of 20 English corpora, totalling more than 17 million words, readable on a variety of systems. All the software tools and manuals needed are also included on its 1999 CD edition. HIT sends, by agreement at cost, the ICAME COLLECTION to users and research institutions. At present there are more than 500 users. It also maintains an information service on the Internet (http://www.hit.uib.no/icame.html).

The ICAME COLLECTION is particularly suitable for comparative studies of English, as it contains American, British, Australian, New Zealand, Indian and East African English corpora of similar sizes. It also includes corpora of early English as well as English of the 1960s and 1990s (LOB and FLOB, Brown and FROWN). Spoken English can be explored by using its more than 2 million transcribed words. Apart from the pioneering Brown and LOB, two corpora of spoken texts in the Collection merit particular attention. One is the London-Lund Corpus (LLC), the first computerized spoken corpus, whose fine prosodic annotation has not been surpassed even today. The other is the Bergen Corpus of London Teenage Language (COLT), the first version of which has been incorporated in BNC. An improved multimedia version with sound will be published late, this year.

Since the establishment of ICAME in 1977, NCCH, later HIT, has also been in charge of the publication of its annual *ICAME Journal* (called *ICAME News* before the advisory board was formed in 1987). The *ICAME Journal* belongs to the ICAME organization and focuses on computers in English Linguistics, whereas the *International Journal of Corpus Linguistics* (*IJCL*), published twice a year since 1996 by John Benjamins Publishing Company, does not belong to ICAME but covers similar grounds, with corpus-based lexicography as its focus. Both the ICAME conference and the *ICAME Journal* are generally regarded as reflecting and leading the trend in English corpora studies.

In addition to its strong support of ICAME-related work, HIT has also successfully conducted a number of corpus or database related projects. One is the Wittgenstein Archives [Bank] (WAB), a fruit of more than ten years of academic research and patient editorial work. WAB comprises 3.6 million words and is published in three electronic versions by the Oxford University Press (OUP). As presented by OUP, the Bergen Electronic Edition is the only CD-ROM series to render instant access to the 20,000 facsimiles and transcriptions of the [Austrian] philosophers hitherto unpublished writings.

As HIT has gained considerable expertise in promoting a unified international text – encoding system, particularly though the experience of WAB, it has become a primary host of the Text Encoding Initiative (TEI) in Europe.

Another project at HIT is the Norwegian Terminological Database (NOT). On March 17, 2000, the NOT database contained 81,687 terms in 30,541 records covering 38 fields. It is still being expanded and linguistically checked. Software programs developed for the NOT database are fairly advanced and have been used in other projects.

NCCH/HIT has been involved with the English-Norwegian Parallel Corpus (ENPC) project in cooperation with Oslo University. The completed ENPC comprises 50 texts in the original and 50 translated equivalents, totalling 2.6 million words intended for comparative studies. Most of the texts in ENPC were scanned with optical scanners at NCCH/HIT, where a program has also been developed to align the original and the translated sentences with an accuracy rate of 95 per cent, which has been adopted by several other countries for their parallel corpora.

The Centre is also a participant of the Henrick Ibsen's Writings project, the most ambitious editorial project in Norway, which started in 1998 and will last for ten years. In this project, HIT will offer its text encoding expertise, choose or develop software for the presentation of texts, and provide the electronic texts and facsimiles of Ibsen's writings that are available at the centre. A previous concordance of Ibsen's plays and poems produced at HIT has laid a good foundation for the present project.

The Norwegian Text Archive, not yet a full-fledged project, has been advocated and partly prepared by HIT. It aims to collect 100 million words of Norwegian text samples. At present it contains 81 million words from newspapers and three million words from novels. A program designed at the centre can collect electronic data automatically and monitor the emergence of new words.

Apart from these, HIT conducts research on applying computer technology to other branches of the humanities, such as dictionary making, library digitization and museums on the web.

It can be concluded from the above briefing that NCCH/HIT has indeed accumulated substantial expertise in corpus linguistics and been in the forefront in this field. The Centre's continuous support has at least in part guaranteed ICAME's success and its ever-increasing international influence. In return, HIT's position as one of the leading European centres in corpus linguistics is enhanced.

Acknowledgements

I am most grateful to the following staff members of the University of Bergen for their time and help: Prof Dr Manfred Thaller, director of HIT until February, 2000; Prof Claus Huitfeldt, project leader of WAB and succeeding director of HIT; Knut Hofland, senior executive officer and information scientist; Kai Innselset, terminologist; Dr Paul Meurer, researcher and software expert; and Prof Anna-Brita Stenström, project leader of the COLT corpus and co-editor of the *ICAME Journal*. My thanks also go to Prof Merja Kytö, co-editor of the *ICAME Journal*, and Sui Changhong for their constructive suggestions. This academic visit is supported by the China Scholarship Committee (CSC).

References

Annual report 1997 and plans for 1998. 1998. Bergen: The HIT Centre.

Annual report 1998 and plans for 1999. 1999. Bergen: The HIT Centre.

ICAME Journal, issues 13 to 23, Bergen: The HIT Centre.

International Journal of Corpus Linguistics. Editorial. Vol.1 (1): iii-x. Amsterdam & Philadelphia: John Benjamins.

Hofland, Knut, Anne Lindebjerg and Jørn Thunestvedt. 1999. *ICAME COLLECTION of English Language Corpora* (Second [Electronic] Edition). Bergen: The HIT Centre.