The Creation of a Spoken Sub-Corpus from the British National Corpus for Comparative Purposes

Phoenix Lam[1] and Richard Forsyth

Abstract

The British National Corpus[2] (henceforth BNC) is one of the most frequently consulted corpora in linguistic research. While the use of this corpus is continuously on the increase, it appears that most BNC-related research work has exploited the corpus in its entirety, i.e. taking the corpus as a whole in analysing specific features or comparing with a different reference corpus. Despite the fact that the BNC is designed with the intention of letting researchers select specific texts to build their own sub-corpora (Burnage and Dunlop, 1993), there are few, if any, studies which generate results from selected texts of the BNC. This paper reports on the process of building a customised contrastive corpus out of the spoken component of the BNC for comparison purposes. In particular, it describes the reasons for making such a customised corpus and the technical issues associated with the selection of texts. It also discusses the topics of balance and representativeness of a corpus (Sinclair, 2005) and the extent to which the BNC represents a wide cross-section of British English from the latter part of the last century.

Although the process outlined in the paper concerns the construction of a sub-corpus of the BNC in order to compare it with a specific reference corpus, the practical issues discussed are applicable in other circumstances involving sub-corpus design from ready-made corpora. In particular, the paper highlights the importance of creating or selecting a suitable corpus for contrastive analyses. It is hoped that the problems addressed in the paper can serve as a source of guidance to future sub-corpus compilers.

[1] *e-mail*: eg.phoenix@polyu.edu.hk

[2] In this paper all references to the British National Corpus refer to the BNC World Edition of the British National Corpus.