BNC: The XML Edition

Lou Burnard[1], Guy Aston[2] and Ylva Berglund[1]

Abstract

A dozen years after its first appearance, the British National Corpus (BNC) remains the most widely available general-purpose fully-annotated English language corpus used by language learners and teachers, corpus linguists, and NLP practitioners. Technology moves on, however, and its SGML format is arguably out of date. Moreover, the many hundreds of users of the corpus have identified many possible improvements in its organization and content. In this paper we present the rationale for, and extent of, the modifications we have been able to make in the latest XML edition of the corpus.

XML is close enough to SGML for migration to be relatively painless and automatic. In producing the new edition, most effort has been devoted to error correction and to enhancement of the encoding. Lacking the resources to add more texts or to carry out a manual proofing and correction of the entire corpus, we have focussed on systematic mark-up errors, and the tidying-up of mistokenisations, misclassifications, and duplicate texts. We will illustrate corrections we could not achieve as well as those which we did. We also describe enhancements to the linguistic annotation: each token is associated with a CLAWS5 pos code (as before); a simpler code taken from a basic 12member tagset; and a root form or lemma. CLAWS multiword units can now be ignored if desired.

Is there still a role for fixed general purpose corpora like the BNC? We believe that the new format will give the corpus a new lease of life by lowering the barrier for those wishing to process its content with general purpose programming tools and integrate it into today's web environments; by improving its quantitative reliability; and by offering wider and more flexible query potentials in conjunction with the XAIRA retrieval software.

[1] Oxford University
e-mail: lou.burnard@oucs.ox.ac.uk
[2] SSLMIT, Bologna