CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model

R. Rossini Favretti, F. Tamburini and C. De Santis
CILTA - University of Bologna - Italy
{rossini,tamburini,desantis}@cilta.unibo.it

A corpus of written Italian – CORIS – has been under construction at the Centre for Theoretical and Applied Linguistics of Bologna University (CILTA) since 1998 and will soon be completed and made available on-line. The project aims at creating a representative and sizeable general reference corpus of contemporary Italian designed to be easily accessible and user-friendly. CORIS contains 80 million running words and will be updated every two years by means of a built-in monitor corpus. It consists of a collection of authentic texts in electronic form chosen by virtue of their representativeness of written Italian.

It is aimed at a broad spectrum of potential users, from Italian language scholars to Italian and foreign students engaged in linguistic analysis based on authentic data and, in a wider prospective, all those interested in intra- and/or interlinguistic analysis. Besides the defined model, a dynamic model (CODIS) has been designed, which allows the selection of subcorpora pertinent to specific research and also the size of every single subcorpus, in order to adapt the corpus structure to different comparative needs. A number of tools have been developed, both for corpus access and for corpus POS tagging and lemmatisation.