# ALLC BULLETIN

Association for Literary and Linguistic Computing
Association de Littérature et de Linguistique Computationelles

Editor: Gordon Dixon

Volume 13 Number 3 1985

## Italian Texts

The major source of Italian texts in machine-readable form, the *Opera del Vocabolario* of the *Accademia della Crusca*, is still inactive, owing to the bureaucratic reorganization. However, a number of less extended projects concerning Italian text processing are now in progress or have been started during this period.

Instead of listing all the new projects which are now underway, we prefer to stress some methodological trends which, at least in my opinion, seem to characterize the situation of research and development regarding Italian texts.

## Sociolinguistics and dialectology

A new project has been announced (NADIR), which aims at creating a very large data base of linguistic data for the south of Italy: not only dialects, but also the so-called 'regional Italian'.

The research will be conducted at different linguistic levels: phonetic, lexical, syntactic, etc. The collection of data will be at the disposal of the community of Italian researchers.

We wish to call attention to the publication of the first (to my knowledge) lemmatized concordance of dialectal texts, which may well become a model in its type. The methodologies for the establishment of the relationship between the dialectal words and their Italian equivalents are extremely sophisticated, and they also rely on a detailed morphosyntactic encoding system.



## Quantitative linguistics

After the publication of the *Lessico di Frequenza dell'Italiano* (Italian Frequency Dictionary), in the early seventies, and the statistical processing of its lexical and morphosyntactic data, very few activities were undertaken in Italy in this field, except for the production of frequency indexes, which always appear with the publication of concordances of literary texts.

We also wish to mention a new project, which may result to be of relevant interest for the methodological development of statistical linguistics.

The corpus of texts in machine-readable form of Italian newspapers of the first half of the 19th century can obviously be divided into subsets according to a number of different criteria and parameters: diacronic aspect, type of newspaper, type of articles, geographical distribution, etc. The research involves different phases, which will benefit from this situation, by applying in an heuristic manner different statistical formulas to the many different experimentally possible subdivisions of the 2,000,000 words of the corpus.

The aims of this project are: for example, to experiment and to compare the power and the characteristics of the different formulas; to gain insights into the relationship between the external production factors (communicative situation, register, etc.) and their linguistic results in frequency distribution; to identify in an inductive manner which linguistic aspects are sensible to which linguistic external factors, etc.

Last but not the least, a frequency dictionary of 19th century Italian will be published, to be compared with the existing frequency dictionaries of the Italian language of the first and second half of the XXth century respectively, from the point of view of the history of the Italian language and culture.

## Towards an Italian linguistic workstation

In our Report on the activities in Italy we describe the studies which are being carried out in the field of machine dictionaries and of linguistic data bases.

A project of the Institute for Computational Linguistics (ILC) relying on the existing Italian machine-based dictionary and the methodology of interactive linguistic data base construction, is aimed at the creation of a linguistic workstation. The intention is to provide Italian researchers with a system in which the machine dictionary, the textual archive, and a number of software modules are integrated. One of the major goals is to utilize knowledge that is implicitly or explicitly embedded at different linguistic levels in the Italian machine dictionary, in order to improve the access of the researchers to the texts. In other words, the researcher will be able to interrogate the texts interactively not only by asking the system to retrieve words or parts of words or sequences of words which he must specify, but also to retrieve 'families of words' which in the dictionaries are connected by some specific relationship: synonymy, semantic field, taxonomy, etc.

*Antonio Zampolli*
*ILC, Pisa*