

**Electronic Archive of the Works by Emilio Gadda:
from Compilers to users of a textual database**

Maria Luigia Ceccotti, Manuela Sassi

ILC-CNR Pisa

TR/10 2000

The project of building the Electronical Archive of Carlo Emilio Gadda's works was born in 1994 when the Garzanti Publisher gave the ILC their texts material published in the "I libri della Spiga" series between 1988 and 1993, already prepared for the photocomposition.

The first version of the Archive, realized in DBT 3.0, CNR patent by Eugenio Picchi, was presented officialy in november 1997 and from that moment was available for study and research purposes.

We will briefly describe the editorial features of this Archive, which now can be consulted in the DBT 4.0 version. This version has a library enriched by new modules and it permits the search not only through the form but also through the lemma.

This Archive has been the starting-point for activities directed to build lexicographical resources to be used either on paper or on electronical sheets. It has been possible to build these resources in three ways:

- by applying simple DBT functions¹;
- by submitting the results to the semi-automatic lematization²;
- by using more than one output of the Archive, transformed into new DBT archives. These archives have been compared by a DBT function and the result has been converted into an electronical sheet which enables us to extract what interests us³.

All this will be done with the help of human control - irreplaceable even if imperfect.

The next step, from the year 1999, has been the planning of an Internet-site on the works of Carlo Emilio Gadda. This is an example of a cultural laboratory containing texts by Gadda, bibliographical data, links and the above mentioned results. The development of this site, to be found in the ILC home page, has been realized applying either the well-known HTML format, or the XML. The XML format is now considered the best instrument for the structuring and the distribution of documents.

This report also attempts to discuss the relation between literature and informatics, which we resume with Giulio Ferroni's words:

"The relation with literature can act to purify language of *pollution*; it can help find hierarchies and distinctions among different spheres of experience. This relation can be the tool for what the american psychiatrist and anthropologist Gregory Bateson (1904-1980) calls the *ecology of the mind*. To make this possible it is necessary to rediscover the pleasure in and capacity for reading. For this to happen one must consider the *literary environment*, the increasing quantity of works and texts, the immense production of documents, which critics or readers can no longer keep up with. The ever growing quantity of literature brings many new problems, which in future probably will have to be tackled by a real *ecology of literature*, which can help to avoid useless and unimportant cultural messages."

Let us try to apply these problems to the information society as a whole, to the labyrinthine web pages in which one can surf free or on payment, and you will agree that the *ecology of literature* is part of the hyperonyme, the *ecology of information*.

Demonstration of the DBT application to the Corpus of Carlo Emilio Gadda's works and the creation of the LR through its functions of search and statistical processing.

The aim of this demonstration is to point out how the use of advanced DBT functions enables us to get different kinds of results from the Gadda Archive. These results, suitably processed, are a library of instruments for the study and analysis of the writer's language; they can be used for literary research and for didactic purposes.

After a first description of how the Corpus has been created and the markup applied, examples of a search will be introduced. We will highlight the search functions which use the morphological engine and the statistic co-occurrences. Their results have been processed to be used and consulted in the author's site at the ILC homepage.

Some of the resources available:

- Apax legomena list, with the title code, arranged in alphabetical order (reverse index) from word-ending to facilitate study of Gadda's lexical production.
- Index Locorum of latin words (in all 5.246 words) and the comparison between these and the wordlist processed from Horace's "Sermones" and "Carmina".

- Concordances, in XML format, of the words (first volume) whose lemma is not recorded in the Italian dictionary Migliorini (1965). The concordances are organized in three lists:
 - a) CONC1 - apax words (frequency=1) with lemma apax,
 - b) CONC2 - words (frequency>1) with lemma apax,
 - c) CONC3 - words (frequency>1) whose lemma is not apax.

The complete wordlist has been analyzed with PiTagger (module of the DBT system) and the result of the lemmatization has been compared with the following dictionaries: DOP-ERI (1969), Garzanti (1987), Zanichelli (1987).

- The comparison between the Garzanti edition of "Quer pasticciaccio brutto di via Merulana" 1957, and the version of the magazine "Letteratura" 1946, published by Alessandro Bonsanti. The results of this comparison are available in the form of a contrastive concordance, where changes are in bold characters.

¹ Apax Legomena list, which is a wordlist subset; index locorum of words with accent, which is a Concordance with zero context of this subset.

² Searching for Gadda's terms. A pre-concordance.

³ Comparison between the two versions of the "Quer pasticciaccio brutto di via Merulana"