The DiaCORIS project: a diachronic corpus of written Italian

C. Onelli*, D. Proietti♣, C. Seidenari♠, F. Tamburini♠

* University of Roma Tre - Italy
♣ University of Modena and Reggio Emilia – Italy
dproietti@unimore.it
♠ University of Bologna - Italy
{c.seidenari, f.tamburini}@cilta.unibo.it

Abstract

The DiaCORIS project aims at the construction of a diachronic corpus comprising written Italian texts produced between 1861 and 1945, extending the structure and the research possibilities of the synchronic 100-million word corpus CORIS/CODIS. A preliminary in depth study has been performed in order to design a representative and well balanced sample of the Italian language over a time period that contains all the main events of contemporary Italian history from the National Unification to the end of the Second World War. The paper describes in detail such design processes as the definition of the main subcorpora and their proportions, the type of documents inserted in each part of the corpus, the document annotation schema and the technological infrastructure designed to manage the corpus access as well as the web interface to corpus data.

1. Introduction

The DiaCORIS project began in March 2004 at CILTA, (Centre for Theoretical and Applied Linguistics, University of Bologna) and is expected to be completed by April 2006. It aims to extend the criteria and structure of the synchronic 100-million word corpus CORIS/CODIS (available online since 2001 and comprising texts from the last two decades of the 20th century) to include written Italian from 1861 to 1945. It is part of a bigger project which aims to produce a diachronic corpus ranging from 1750 to 1945, one of the various activities supported by a publicly financed national project – FIRB.

DiaCORIS introduces a diachronic perspective to analysis of the Italian language, providing scholars with a powerful and flexible tool to analyse linguistic phenomena over a long period of time which covers all the main events of contemporary Italian history from the National Unification to the end of the Second World War.

These characteristics make DiaCORIS unique in the not very rich panorama of existing Italian corpora, or electronic collections of texts, and as a reference for historical study of Italian language, literature and culture.

Two corpora of different dimension and design are freely available on–line for research on old Italian. The first one is the data base of OVI (*Opera del Vocabolario Italiano*), a rich corpus including all the literary and non literary texts in prose and poetry written in early/old Italian from the beginning of the XIII century to 1375 (the year of G. Boccaccio's death, the last of the great authors writing in the XIV century: Dante, Petrarca and Boccaccio himself, constitute the so called "tre corone" = " three crowns" of Italian language and literature). The very large collection of texts of the OVI data base (about 19 million words) is accessible through query tools that allow the automatic generation of concordances and word lists.

The second one, BIVIO, has been realized thanks to a collaboration between the Istituto di Studi sul Rinascimento and the Centro di Ricerche Informatiche per le Discipline Umanistiche (Signum) c/o the Scuola Normale Superiore di Pisa. BIVIO (acronym of *Biblioteca Virtuale On-Line*), is one of the various textual resources amongst dictionaries, repertories, and literary works mainly relating to the history of Italian renaissance fine arts accessible on-line in the Cribecu site (Centro di Ricerche Informatiche per le Discipline Umanistiche c/o Scuola Normale Superiore di Pisa). The corpus has gathered about 200 literary and essayistic works (some of these rare but of great importance from a historical and cultural point of view) by about 60 authors of the XV-XVII centuries. The texts collected mostly relate to philosophy and history of fine arts. A powerful query program allows the user to retrieve many kind of information (frequencies, concordances, links between texts and images) either on a single text, on a subcorpus or on the whole BIVIO corpus.

A third resource is available only on CDRom: LIZ (*Letteratura italiana Zanichelli*), which aims to be a reference corpus for research on Italian literature, contains literary texts (1000 works in poetry or prose from the XIII to the XX-century-writer L. Pirandello), with a larger selection of writings for the XV-XVII centuries. As in BIVIO, the query program of LIZ (DBT, *Data Base Testuale*, designed by the ILC-CNR) is designed to be able to retrieve data information from a single text, from a subcorpus as designed by the user, or from the whole corpus.

Queries only for single texts are possible for other collections available on the net such as the Italian section of the Biblioteca Digitale IntraText powered by Èulogos.com (2575 texts, in mainly religious, theological and moral texts). The texts collected in the Biblioteca Digitale IntraText are presented and available as lexical hypertexts: every word is linked with the frequency data and a brief context for every occurrence. In the lexical hypertextualisation of the texts, however, stop words are not considered: that makes these hypertexts hard to use for linguistic researches.

Large collections of free downloadable Italian texts are available in no profit association websites such as the world-wide extended and known Project Gutenberg which is the first producer of free e-books on the Internet (about 1000 texts are collected in the Italian section) and the Italian association Liber liber - Progetto Manuzio (about 1200 literary and non literary text). Both Project

Gutenberg and Progetto Manuzio are mere text archives and they were not intended or realized as textual corpora.

The same holds true for collective works on CDRom such as *Letteratura italiana* and *Storia d'Italia* Einaudi.

Against this background, DiaCORIS stands out as a collection of texts organically conceived on the model of CORIS/CODIS. DiaCORIS is a representative collection of all the most important textual types and genres of the Italian language in XIX and XX century culture (described more specifically below), and it is also a resource which has been carefully organized to be a reliable and powerful tool for diachronic linguistic research in this domain.

2. Corpus design and representativeness

DiaCORIS includes written Italian texts produced between 1861 and 1945; all the texts collected belong to a level of Italian language which, using the designations determined by Nencioni (1983), can be described as "written-written". To fully exploit the diachronic aspect of the text collection the time span was split into three periods as widely acceptable as possible from a historical and socio-cultural point of view: "After National Unification", "The Liberal Period", "Fascism". Three homogeneous subcorpora correspond to the periods proposed, each made up of 5 million words; the total size will thus be 15 million words.

As DiaCORIS is also intended as a supplementary resource for CORIS/CODIS, the question of direct comparability of the two corpora was addressed from the beginning. The texts selected for each of the three DiaCORIS subcorpora are divided, as in the CORIS/CODIS corpus, into macro-varieties on the basis of textual and external (pragmatic) characteristics. This unambiguous categorisation makes it easily comparable to other corpora and reduces researcher subjectivity to a minimum.

SUBCORPUS	PROP.
PRESS	38%
FICTION	25%
ACADEMIC PROSE	12%
LEGAL-ADMINISTRATIVE PROSE	10%
MISCELLANEA	10%
EPHEMERA	5%

Table 1: CORIS/CODIS structure.

The main varieties of written Italian identified as subcorpora of CORIS/CODIS (as shown in Table 1) were essentially reproduced with some minor modifications. The "Ephemera" section of CORIS, consisting mainly of heterogeneous texts characterised by a short time circulation, could not be included given that the widespread use of such texts became a significant phenomenon in Italy only in the second half of the 20th century. Moreover, compared with the corresponding CORIS section "Academic Prose", the DiaCORIS "Essayistic Prose" contains a wider but still comparable selection, including study and critique texts from different disciplines (from literary, musical and artistic criticism to political pamphlets, to the historical and scientific prose). This is an extremely well-articulated kind of text, whose decisive importance in the process of development of modern Italian society and in the correlative definition of a modern common use of Italian language (the so called "italiano dell'uso medio", Sabatini 1985) is universally recognized but has not yet been deeply studied (Proietti 2004, p. 347).

As Table 2 shows, a variable structure was preferred for the subcorpora. The proportion of almost all the sections changes across the subcorpora in order to take into account their varying contribution as representative samples of the evolving Italian language. The increasing size of the "Press" section, for instance, testifies to the increasing importance within Italian society during the period under consideration of texts from the mass media owing to the growth in both production and readership of newspapers and related press.

SECTION	SUBCORPUS	PROP.
1861-1900	PRESS	15%
After	FICTION	30%
National	ESSAYISTIC PROSE	30%
Unification	LEGAL-ADMINIST. PROSE	10%
	MISCELLANEA	15%
1901-1922	PRESS	25%
The	FICTION	25%
Liberal	ESSAYISTIC PROSE	25%
Period	LEGAL-ADMINIST. PROSE	10%
	MISCELLANEA	15%
1923-1945	PRESS	30%
Fascism	FICTION	25%
	ESSAYISTIC PROSE	25%
	LEGAL-ADMINIST. PROSE	10%
	MISCELLANEA	10%

Table 2: DiaCORIS structure.

Having defined these macro-varieties, we decided to make a further division into subsections and sub-subsections. Again, these were based on external parameters which would allow the data to be contextualised. For example, it was clear that a sampling of the "Press" texts would have little value without further refinement in order to take into account the socio-cultural factors influencing the respective periods of each subcorpus. In all the three subcorpora, therefore, the reference to the above-mentioned parameters led to the division of "Press" into two more subsections, namely "Newspapers" and "Periodicals", each including texts from various geographic areas (e.g.: from newspapers of different cities, such as the *Corriere della sera* [Milano], *La Stampa* [Torino], *La Nazione* [Firenze], *Il Mattino* [Napoli], *Il Giornale di Sicilia* [Palermo]), or characterized by different ideological and political attitudes (e.g.: articles from republican, socialist and catholic periodicals in the Liberal period; a rich selection of articles and essays from the most relevant of the numerous periodicals of the first two decades of the XX century). Then, in the third subcorpus ("Fascism") particular attention has been given to obtaining representative texts for three balanced sub-sections: "Generic/Regime Press", "Underground/Opposition Press" and "Catholic Press".

It is important to point out the significance of the collection and sampling of the source data from newspaper and periodicals which were vitally important for the development of Italian society, culture and language. Rare articles and texts by different authors and from different cultural and political attitudes, which can

be hard to find in Italian libraries, will be available on the net for the first time as a comprehensive diachronic corpus. For example: some popular magazines of the second half of XIX century chosen among those edited by Dina Bertoni Iovine; nationalist newspapers like *L'idea nazionale*, Mussolini's fascist *Popolo d'Italia*, the racist magazine *La Difesa della razza*, antifascist underground newspapers and magazines during the second world war.

Finally, we decided to only include normative texts (laws and codices) in the legal-administrative prose section, excluding the forensic-jurisprudential and administrative writings which are more related to the context of their production and which felt the effects of the drastic institutional and political upheaval in the 1861-1945 period more heavily. The normative texts, on the other end, due to the fact that they are ruled and characterized by a "binding" use of the language (Sabatini 1990; 2001), are more stable, regular and therefore more easily comparable over a long period (1861-1945).

With the aim of taking into account as wide a range as possible of textual types and genres, within the Miscellanea section we pursued and collected works which are quite different from one another in terms of their destination or their cultural and linguistic peculiarities. Besides the more popular and widespread novels and texts for children (from Collodi's *Pinocchio* to Salgari, to Vamba's *Gianburrasca* and Yambo's *Ciuffettino*, till the fascist scholastic book *Balilla Vittorio* by R. Forges Davanzati) we decided to include serial stories such as the novels of C. Invernizio or comic novels like the humorous *Come ti erudisco il pupo* by L. Lucatelli. Furthermore, to survey the role and the function of translators from foreign languages in shaping the "italiano dell'uso medio" (everyday Italian), translations by popular writers were collected, such as those by F. Verdinois (from Russian and French), by C. Pavolini (*Le avventure del barone di Münchhausen* by R. E. Raspe), by C. Sbarbaro (*À rebours*, by J.-K. Huysmans) or, those by anonymous translators at the beginning of XX century, such as translations of the well known novels by L. Carroll. Moreover, we put together private works or documents (such as the diary by F. Tozzi or the *Lettere dal carcere* by A. Gramsci) and public writings such the official acts of the short-lived "Reggenza del Carnaro (Carnaro Regency)" by G. D'Annunzio or the encyclical of the Roman popes (from Leo XIII to Pius XII).

Within the "Saggistica" section we tried to give examples of all the various textual types or genres and of the different disciplines and cultural domains, again paying particular attention to the "uso medio" from a linguistic and cultural point of view. We therefore gathered a number of essays and manuals in a collection of representative samples of works such as the brilliant etiquette-book *La gente per bene* by the Marchesa Colombi or the popular set of "Fisiologie" by the scientist P. Mantegazza; from among the socio-political pamphlets we chose the text the nationalist-socialist speeches by G. Pascoli, portions from *Quaderni del carcere* by A. Gramsci and works of various kinds by S. Sighele, G. Salvemini, P.Gobetti, G. Dorso, C. Levi. Furthermore, we selected treatises and historical essays and collected texts such as the *Elementi di scienza politica* by G. Mosca, *Il metodo della pedagogia scientifica* by M. Montessori, or the *Compendio di sociologia* by V. Pareto and *La terra e l'imposta* by L. Einaudi. Among the historical studies we collected works such as the *Storia d'Europa* by B. Croce and the *Pensiero italiano del Rinascimento* by G. Gentile; philosophical and literary essays are represented by masterworks like the *Estetica* by B. Croce and *L'umorismo* by L. Pirandello. From among the writings of literary theory we chose the *Manifesto tecnico* of Futurism by F.T. Marinetti and the *Esame di coscienza di un letterato* by E. Serra; as representative examples of a textual kind between the essay and and literary prose (in particular, the so called "prosa d'arte") we chose works such as *Il fanciullino* by G. Pascoli, *L'ora topica di C. Dossi* by G. P. Lucini, *Il sole a picco* by V. Cardarelli and *Cristo s'è fermato a Eboli* by C. Levi.

Finally, a great effort was made in taking into account all the textual typologies and the entire range of linguistic varieties and stylistic registers for the Fiction section. Here, beside works of naturalistic inheritance or (neo)realistic inspiration (like *Il marchese di Roccaverdina* by L. Capuana, *Il mulino del Po* by R. Bacchelli, till *Gente in Aspromonte* by C. Alvaro), we collected novels of a psychological and introspective nature such as *Il fu Mattia Pascal* by L. Pirandello, *La coscienza di Zeno* by I. Svevo, the principal novels by F. Tozzi (*Con gli occhi chiusi*; *Il podere*; *Tre croci*) and *Le sorelle Materassi* by A. Palazzeschi. Novels of popular destination and diffusion (such as those by G. Deledda) are, on the other hand, juxtaposed near examples of linguistic expressionism (*Fosca* by I.U. Tarchetti; *Racconto italiano di ignoto del Novecento* by C.E. Gadda; *Casa "La vita"* by A. Savinio) or samples of lyric or calligraphic description (*Il mio Carso* by S. Slataper; *Un anno sull'Altipiano* by E.Lussu).

3. Corpus access and annotations

The documents included in each of the three DiaCORIS subcorpora have been enriched with metadata information for the principal document data such as "author", "date of publication", "publisher or serial", "genre", etc., using XML schema. This kind of annotation allows the researcher to refine corpus queries and gather highly specialised information.

The requirement to manage structured documents as well as performing complex queries on annotated data in a quick and easy way led us to choose the IMS-CWB package (Christ, 1994) as corpus administration infrastructure. It offers partial support for document structural annotations following a simplified XML mark-up schema as well as a fast and powerful query language.

Every text included in DiaCORIS has been annotated following the schema outlined in table 3.

```
<text id="filename.xml" lingua="Italian"
      titolo="document title" autore="author"
      ed_test="editor" anno="publication year"
      sez="section" subc="subcorpus">
word1  pos_tag1
word2  pos_tag2
…
</text>
```

Table 3: DiaCORIS document structure.

A web interface has been built on top of the IMS-CWB text retrieval engine allowing fast selection of the

various querying options as well as query restriction to a specific corpus section and/or subcorpus (see figure 1).

The corpus interface resemble the one built for CORIS/CODIS access, maintaining similar fields and options though the use of different underlying corpus retrieval engines. Two different querying styles have been proposed: the first is identical to the CORIS/CODIS style, thus it is translated to conform to the IMS/CWB engine query language while the second reproduces the IMS/CWB language options exactly.

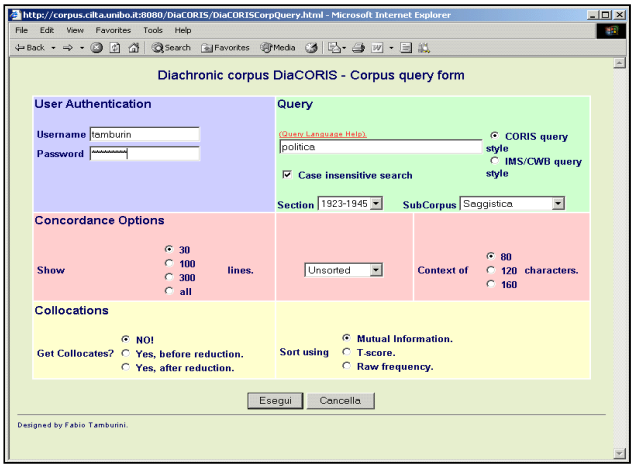

Figure 1: DiaCORIS web interface to query language.

Once the concordances of a given term have been retrieved (see figure 2) it is possible to extract the wider context and all the document information from which a single concordance derives (figure 3).

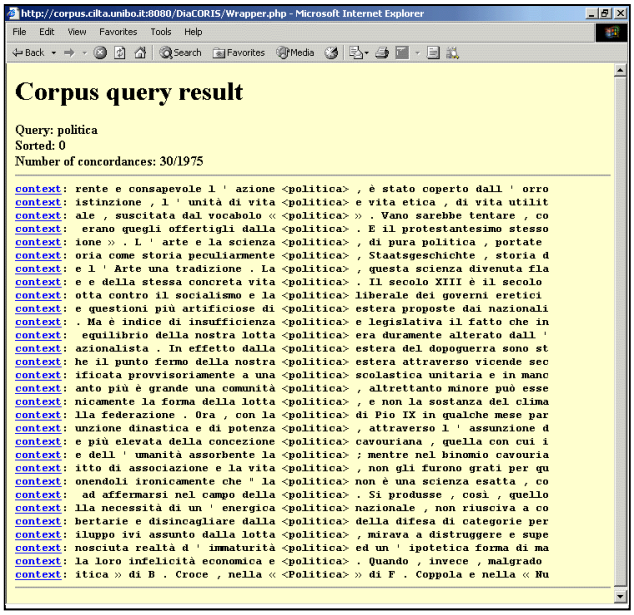

Figure 2: Corpus query result.

DiaCORIS will be freely available for research purposes at the end of the project.

Other types of annotation, primarily part-of-speech tagging, will be introduced to further enrich the access possibilities to the corpus data before the end of the project.

4. References

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system, In *Proc. COMPLEX'94*, Budapest.

Nencioni G. (1983), Parlato-parlato, parlato-scritto, parlato-recitato. In Id., Di scritto e di parlato. *Discorsi linguistici*, Bologna: Zanichelli, 126-179.

Proietti D. (2004), Saggio. In Le Muse. *Dizionario enciclopedico*, X, Novara: De Agostini, 342-347.

Rossini Favretti R., Tamburini F., De Santis C. (2002), CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. In *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, Wilson, A., Rayson, P. and McEnery, T. (eds.), Munich: Lincom-Europa.

Sabatini (1985), L'"italiano dell'uso medio": una realtà tra le varietà linguistiche italiane. In *Gesprochenes Italienisch in Geschichte und Gegenwart*, Holtus, G., Radtke, E. (eds.), Tübingen: Narr, 154-184.

Sabatini (1990), Analisi del linguaggio giuridico. Il testo normativo in una tipologia generale dei testi. In ISLE – Scuola di scienza e tecnica della legislazione. *Corso di studi superiori legislativi, 1988-1989*, D'Antonio, M. (ed.), Padova: CEDAM, 675-724.

Sabatini (2001), I tipi di testo e la "rigidità" del testo normativo giuridico. In *La scrittura professionale*, S. Covino (ed.), Firenze: Olschki, 97-105.

Tamburini F. (2002), A dynamic model for reference corpora structure definition. In *Proc. LREC2002*, Las Palmas, Spain, 1847-1850.

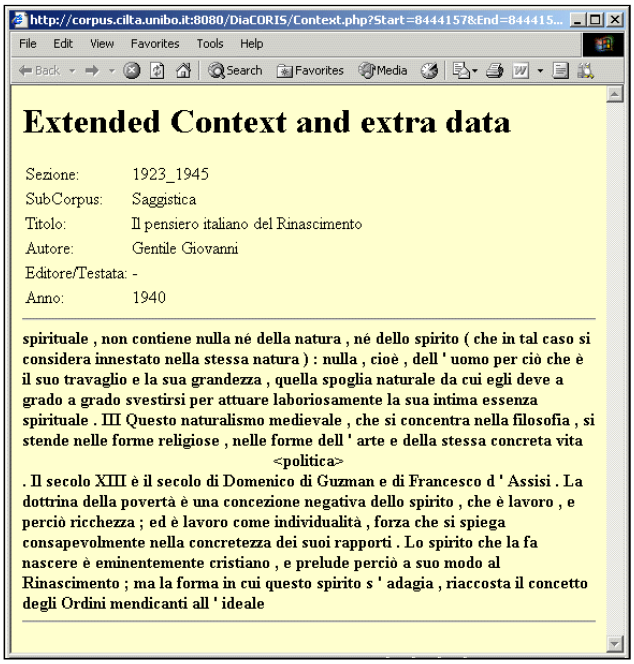

Figure 3: Extended context and document data for a given concordance.