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

Abstract

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.

1. Introduction

The aim of this paper is to describe the most important steps in the design and construction of CORIS - *COrpus di Italiano Scritto* - an electronically-based reference corpus of contemporary written Italian containing 80 million running words.

The project is the result of research carried out at the University of Bologna. This was possible thanks to technological development, previous experience[1] and the identification of design criteria that preceded the planning and construction phases.

To describe the realisation of CORIS briefly, the principle phases may be indicated as follows:

- CORIS Corpus design
    - Corpus typology
    - Corpus size
    - Representativeness
- Design of source text framework
    - Text typology
    - Text unit size
    - Definition of selection criteria
- Corpus structure
    - Subcorpora definition
    - Subcorpora-to-subcorpora ratio
    - Definition of sampling criteria
- Text acquisition and copyright
- CODIS - a dynamic model
- Corpus access tools
- Part-of-speech tagging and lemmatisation

The design of the corpus will be discussed focusing on some of the main issues such as representativeness and comparability. Comparability can be seen as the main reason for the construction of CODIS - COrpus Dinamico di

[1] See in this connection the construction and analysis of BOLC – BOnonia Legal Corpus – a multilingual comparable legal corpus being developed at CILTA since 1997, in which John Sinclair played a crucial role as consultant. Rossini Favretti 1998, 2000.

Italiano Scritto – an alternative corpus structure which is intended to be dynamically adapted to meet the specific requirements of individual researchers.

2.1. CORIS Corpus design

In order to design and construct CORIS, some preliminary choices were necessary to lay the foundations for successive stages. First of all we defined the aim of the project , and the type of corpus it was intended to create. From the very beginning, the purpose of the project was to construct a general corpus, as defined by the Brown Corpus, one of the first electronic corpora. Just as the Brown Corpus was referred to as "a standard sample of present-day English for use with digital computers", so too the aim of CORIS, at the design stage, was to create a collection of texts in electronic format representing, in the widest sense, present-day Italian. The identification of this aim provided a solution to one of the first issues which arose in the planning of the corpus, the choice between synchronic and diachronic dimensions. It was decided to select texts synchronically in order to permit a generalised description of commonly used Italian.

The choice between written or spoken language gave rise to greater problems. Having taken into account various possibilities, bearing in mind the obvious advantages of having a corpus consisting of both written and spoken texts, it was decided to give priority to written texts at this stage of research. The decision was based both on external and internal criteria. First of all, it was influenced by the general panorama of Italian linguistics and the position that the corpus would occupy alongside works such as the *Lessico di frequenza dell'Italiano Parlato* (LIP, 1993), *Lessico di frequenza della lingua italiana contemporanea* (LIF,1972), *Vocabolario elettronico elettronico della lingua italiana* (VELI, 1989) and (*Letteratura Italiana Zanichelli in cd-rom* (LIZ, 1993[1],1995[2], 1997[3]) to name just the most significant. We should also mention the corpus of Italian developed at the ILC by the Pisa CNR as part of the PAROLE project (1996-1998). Secondly, in the light of transformations in communication by new technologies, it was preferred not to pose the problem of the relationship between the language traditionally considered as standard spoken Italian and its technological ramifications via telephone, radio, television and/or computer technology.

For these reasons, it was decided to develop a synchronic corpus of written language, using texts, roughly speaking, from the 1980s and 1990s, with a somewhat longer timescale as far as narrative is concerned. They belong to an Italian language which, using the criteria proposed by Nencioni (1983), can be described as 'written-written'.

The definition of the size of CORIS was more problematic. A study of presently available corpora clearly revealed that it was not possible to make reference to any standard size. The rapid and widespread development which, especially in recent years, has characterised both the low-cost availability of hardware and the production of more efficient and user-friendly software has radically transformed the criteria for the creation of the most recent corpora compared with those of the first or second generation.

While the criteria on which first generation corpora, such as the Brown Corpus, appear to have been mainly influenced by the potentiality of information technology, present-day technology no longer sets any limits to the choices of the researcher, who can extend a corpus to include the varieties held to be relevant to his/her analysis and, then, make a suitable selection of the varieties of representative texts. Developments in information technology over the past years, the present speed of the processing of material and the low cost of mass storage units mean that it is possible to create corpora consisting of hundreds of millions of words, such as the British National Corpus and the Bank of English. Especially it would seem that, as far as written language is concerned, the standard of one million words has given way to a standard of one hundred million.

However, any generalisation is debatable, as is any definition of a set limit. The Brown Corpus (1967), with one million words, 500 written text samples of 2000 words each, representing in equal measure the main text types, is still considered by many scholars to be a valid model. One of the most recent English language corpora, the Longman Spoken and Written English Corpus - LSWE Corpus - created by Biber, Johnson, Leech, Conrad and Finegan, consists of about 40,000,000 words and contains 37,244 texts. It is stated that these texts vary in length according to register.

A further aspect to be considered in the definition of a corpus is that of the introduction of monitor corpora. These provide for constant updating by means of the periodic introduction of data carried out by a collection of filters, on the basis of a selection of new and existing data. The configuration of the monitor corpus means that the aspects of determinacy and permanence which were defining characteristics of the size of a corpus over the past decades are no longer valid. The corpus takes on a dynamic configuration, which seems more relevant and advantageous if we consider that today, with the possibilities provided by the development of new technology and memory, it is no longer necessary to go to the trouble of selecting texts. It seems to be possible to manage a corpus the principal components of which are delimited and, at the same time, a monitor corpus which is open and able to record innovations and modifications in current usage. This combination makes it possible to access a corpus which is available in a finite form - either on-line or on CD-Rom - and can be updated by means of the monitor as well as by the introduction of supplementary subcorpora representing further varieties.

It was therefore decided to proceed with the planning of a corpus the size of which, though set as "large", was not predetermined but related to the choice of linguistic varieties thought to be representative and, as such, set as an intermediate research goal following the compilation of a pilot corpus.

The definition of representativeness is a crucial point in the creation of a corpus, but it is one of the most controversial aspects among specialists, especially as regards the ambiguity inherent in its use due to the intermingling of quantitative and qualitative connotations. While for some scholars the extension of corpora to include hundreds of

millions of words might make up for a slight differentiation in the varieties represented, for others a wide differentiation in varieties is seen as an essential condition for any act of generalisation.

Even in the first phase of our research the problem of representativeness did not, in our opinion, disappear with the possibility of enlarging the corpus; indeed, it was accentuated. In spite of the increase in size to hundreds of millions of words, each corpus represents a limited sample of language in use. An operation of sampling, however extensive it may be, inevitably turns out to be simplified in the light of the complexity of the phenomenon under examination. Even building random selections into the corpus construction, it seemed to us that in the transition from the sample to the generalisation, certain degrees of approximation should be provided for, thus allowing maximum flexibility and dynamics in the proposed model.

In the light of problems of epistemological nature encountered in the planning of a corpus which could be defined as being representative of a language or the state of a language, it was decided to proceed recognising the limits inherent in the project and identifying parameters which might counterbalance those limits. Some criteria of identification for the parameters of reference were thus defined which permitted the creation of a collection of sub-corpora which included the main varieties of written Italian, represented and appropriately balanced. At the same time, it appeared possible to construct a dynamic and adaptive model which would satisfy the needs and working hypotheses of different scholars while still respecting the criteria of corpus construction.

2.2. Design of source text framework

In the context of corpus linguistics, one of the basic criteria accepted by all projects and studies is the fact that selected texts must be authentic and commonly used in social interaction. However there is no consensus as to whether to enter texts in their entirety or in fragments that are considered to be representative. This is indeed a crucial issue and was the object of considerable reflection at the planning phase. As we have seen, in the first corpora, such as Brown, standardised sampling was applied. Uniformity of text size was one of the basic construction principles. If there was disagreement, this focused upon the size of the samples. In the designing of the construction model, considering the present conditions created by software programs, the problem was not so much that of defining sample size but rather of the choice to be made between texts and texts fragments.

The first inevitably leads to a lack of standardisation of text samples. It is rarely the case that several texts, whether they be journalistic, narrative or scientific, contain the same number of words. The second, on the other hand, may lead to a stronger influence of the researcher's subjective judgment and implies that the selected sequence is taken out of context. This could mean that the larger size actually invalidates the representativeness of the corpus. It was therefore decided that, where possible, the entire text would be entered, rather than standardising sample size.

A later step was the definition of linguistic varieties used to create the corpus. These are considered as a collection of documents identifiable on the basis of both external and internal features, in which the peculiarity of the single variety fades away in comparison to the mass of data. This constituted one of the most important points. Although the corpus included specialist areas, such as legal, scientific and bureaucratic-administrative language, an attempt was made to bring together not so much a collection of specialist texts as a variety of types which, according to our investigations, can be placed within a continuum, overlapping and integrating one and another.

When defining the selection and construction criteria, reference was made to both external and internal parameters in order to reduce the researcher's choice to a minimum. Furthermore, considering the context of CORIS as well as the wide availability of existing and planned corpora, a further criterion was introduced, that of "comparability", in order to offer scholars the possibility of interlinguistic comparison of corpora.

2.3. Corpus structure

In order to define the initial construction phase of the corpus, what we would describe as criteria of external textual features and comparability were of prime importance. These led to the identification of the initial construction phase - provided by the sub-corpora - in which it was possible to refer to some macro-varieties identified on the basis of external appearance or the material elements of the text, extremely clear in their characterisation and easily comparable. The subjective choices of the researcher would thus be reduced to a minimum.

As the distinction between "published" and "unpublished" texts was considered to be too simple, various kinds of publications from the "press", "narrative" from various types of volumes and essays identified as miscellaneous were then selected, and various hand written, printed and above all electronic texts were grouped together in a section under the heading of "ephemera" due to their transitory nature.

Having defined these macro-varieties, it was thought necessary to apply a second level of articulation - based on the sections which could be divided into subsections - which, again using external parameters as a basis, still enabled collected data to be contextualised. For example, it was clear that a sampling of the "press" population could not be undertaken except on the basis of a second phase connected to the socio-cultural reality of the nation. This was considered to be a fundamental point in order to arrive at a definition of a population's components, albeit with some degree of approximation.

The reference to the above-mentioned parameters led to the following structure:

Subcorpus	Sections	Subsections
PRESS	newspapers, periodic, supplement	national, local specialist, non-specialist connotated, non-connotated
FICTION	novels, short stories	Italian, foreign for adults, for children crime, adventure, science fiction, women's literature
ACADEMIC PROSE	human sciences, natural sciences, physics, experimental sciences	books, reviews scientific, popular history, philosophy, arts, literary criticism, law, economy, biology, etc.
LEGAL AND ADMINISTRATIVE PROSE	books, reviews	legal, bureaucratic, administrative
MISCELLANEA	books, reviews	books on religion, travel, cookery, hobbies, etc.
EPHEMERA	letters, leaflets, instructions	private, public printed form, electronic form

Table 1: CORIS corpus structure.

Having defined the selection criteria, the next step was the planning of the sub-corpora, first examining the size they should have and the ratio between the size of the various subcorpora and sections.

An initial idea was to consider the possibility of working on the basis of a randomised selection and to correlate the dimensions of each subgroup of texts to the number, albeit approximate, of the recipients of a given text. The application of quantitative parameters - such as circulation and distribution - proved to be too limiting in comparison with qualitative parameters such as time and type of text use or level of cognitive attention. So despite the difficulties involved in the introduction of qualitative (hence non-measurable) parameters, it was our opinion that merely quantitative data were not sufficiently significant and that they should be integrated, as far as the percentage ratios between sub-corpora and sections was concerned, with qualitative variables, lest any one variety should be overestimated. This choice of procedure was corroborated by an in-depth analysis for 1997:

PRESS (data derived from FIEG, La stampa in Italia 1995-1998, Milano, 1999)		BOOKS (data derived from AIE, La produzione libraria italiana del 1997, Milano, 1999)	
Newspapers	2,955,501,360	Fiction	119,100,000
Weekly magazines	730,364,544	Non-fiction	179,400,000
Monthly magazines	194,607,972		
TOTAL	3,880,473,876	TOTAL	298,500,000

Table 2: Quantitative data used for corpus design.

The ratio of 12:1 established, more or less, between texts from the mass media and books could not be accepted as being reproducible in the samples. On the other hand, it appeared to be too significant to ignore, even bearing in mind the comparability of the corpus under construction. Within the ratio allowed by the sales volumes, which, on the basis of the data, is represented as an interval, it was decided to set the ratio between the different areas of circulation as the smallest allowed value in order not to penalise certain textual varieties, such as letters.

Having selected a wide range of linguistic varieties, texts for the entry of the single sub-corpora were prepared and, in order to comply with the criterion of representativeness, the documents were randomized within each sub-corpus. Having defined this objective corpus framework, the following macro-varieties were considered:

PRESS	30 million words
FICTION	20 million words
ACADEMIC PROSE	10 million words
LEGAL AND ADMINISTRATIVE PROSE	8 million words
MISCELLANEA	8 million words
EPHEMERA	4 million words

Table 3: Sizes of the principal macro-varieties in CORIS.

Therefore, the corpus of written Italian - CORIS - was built along general lines as:

> *a collection of texts which are authentic, commonly occurring, in electronic format, chosen as representative of present-day Italian*

and in terms of size as:

> *a general corpus consisting of 80 million words updated every two years by means of a monitor corpus*

CORIS was designed and built as a general reference corpus for the analysis of written Italian and will be on-line in the middle of 2001.

3. Text acquisition and copyright

After the decisions on the design of corpus were taken and the criteria for selection of texts were defined, two practical matters had to be faced: the acquisition of authentic texts in electronic form and permission to include the texts in the corpus and to quote extracts from them in various publications.

Focusing on the first problem, it may be worth considering how the situation has been evolving during the last 10 years. Sinclair (1991) describes three methods of text input, each suitable for a different class of material:

a. adaptation of material already in electronic form (for example, newpapers on CD-Rom)
b. conversion by optical scanning (which is considered the best option for books printed by conventional methods)
c. conversion by keyboarding (especially for handwritten material).

At the present time, since keyboarding and conversion by scanning have become too slow and expensive for the aim of constructing and updating a large corpus of written language, the adaptation of electronic texts – downloaded off-line (from CD-Rom) or on-line (from the Internet) – is becoming established as the sole source of contents.

Our project made an extensive use of this method for the acquisition of Press, Legal and Administrative Prose and other documents included in the Miscellanea subcorpus. Moreover, materials such as leaflets, instructions or letters (included in the Ephemera subcorpus), which were traditionally handwritten, could be taken directly from Web pages or e-mail.

Also for printed books (collected in Fiction, Academic Prose and Miscellanea subcorpora), it was often possible to bypass the labour of putting material in electronic form: in modern methods of publishing there is always an electronic stage, and the co-operation with the most enlightened publishers made it possible for us to acquire many texts in electronic form. In these cases it was sometimes necessary to convert the original format of the text (PDF or other printing formats) into ASCII plain text and to standardize the accents (with ISO-Latin 1 encoding).

In other cases, it was the "replicator technology" of the Internet – as Michael Hart, the beginner of Project Gutenberg, defined it referring to Walter Benjamin (1966) – that provided resources suitable for our purpose. Archives of electronic texts, which began on the Internet in the 1970s following Borges' dream of a "universal library", have increased rapidly in Italy during the last years (see for example LiberLiber by Progetto Manuzio), making available copies of books (usually in ASCII text, the simplest and easiest to use format) that are freely accessible to users (Calvo *et al*. 2000, De Santis 2000, Spina 1997).

However, pursuant to Italian and EU law, the fact that documents can be downloaded does not imply that they can be reproduced and distributed without copyright permission (Bertola 1999, Revelli 1997). Hence the need for a number of requests to obtain permissions from the copyrights holders, especially for printed texts to be digitally reproduced.

Shifting our attention on this problem, we should notice how Sinclair's (1991) words are still relevant: "the labour of keeping a large corpus in good legal health is enormous". A lot of paperwork was necessary to explain to publishers why the texts were needed (for academic research purposes, implying a non-commercial use) and what safeguards there could be against exploitation and piracy (the text would be compressed and indexed, it would be consulted only in KWIC format, extracting small quotations from the original, and would not be given to a third party). However our efforts did not always bear fruit.

Intellectual property laws have not been updated in order to take account of the technological changes of the last decades[2] and the normal authorial contract does not make provision for the use of a text in electronic form.

Furthermore, the distinction (Vitiello 1997)made in Italian legislation on "diritto d'autore" between *material rights* (which last throughout the author's life and for 70 years after death; they can be sold and cover copying, reproducing etc.) and *moral rights* (which are everlasting: they cannot be sold and include the paternity right and the integrity right) multiplied the number of requests to forward. The first had to be addressed to the copyright owner (usually the publisher), asking the permission to made a temporary copy of the text, and the second to the author or the assignees in order to safeguard the moral rights besides the material interests[3]. These are the reasons why the negotiations were

[2] According to Samuelson (1991), the reasons for the inadequacy of the intellectual property law can be traced back to certain distinctive qualities of the digital medium: ease of replication, ease of transmission and multiple use, plasticity of digital media, equivalence of works in digital form, compactness of works in digital form, non-linearity.

[3] Moral rights become particularly important in this context since the original text will be modified: it will be entered in a corpus and made accessible by the concordance program in a way very different from the intentions of the author. See Chimienti 1999 for further details on the problem of copyright in the creation of databases and multimedia works.

particularly long and difficult. Nor does the Italian system encourages private agreements, made in order to directly preserve the interests of the parties (Marandola, 1996).

Of course there are occasions when copyright does not apply and materials may be used without a licence: in general this involves the reproduction or quotation of extracts for research, study or criticism. The aim of corpus linguistics could be easily included in this range of fair uses, while waiting for special provisions. However, it seems more and more urgent to solve this problem at an international level, so to ensure a proper balance between the need to protect the interests of authors and publishers and the need to allow access to users.

4. CODIS - a dynamic model

Considering the vital role which will be played by the comparability of a reference corpus, it seemed important to provide for the possibility of creating an alternative corpus structure which would make it adaptable to the needs of different researchers. As shown in table 1 the structure and proportions, among the different subcorpora that composes some corpora, varies considerably.

CORPUS	COMPOSITION	
BNC - 90Mw – English Written section	Books Press Miscellanea	52.5 Mw - 58.6% 27.8 Mw - 31% 7.4 Mw - 8.3%
LSWE - 28Mw – English Written section	Fiction News Academic Prose General Prose	5 Mw - 17.8% 10.6 Mw - 37.7% 5.3 Mw - 19% 6.9 Mw - 24.6%
The Oslo Corpus - 22.3 Mw – Norwegian	Fiction Newspaper/Magazine Factual prose	3.8 Mw - 17% 10.6 Mw - 47.5% 7.8 Mw - 35%
Corpus de Referência do Português Contemporâneo (CRPC) - 92 Mw – Portuguese Written section	Newspaper Books Periodical Decisions of Supreme Court of Justice Miscellanea Leaflets Correspondence	55 Mw - 60.8% 20.5 Mw - 22.6% 7 Mw - 7.7% 1.8 Mw - 2% 3.9 Mw - 4.3% 0.3 Mw - 0.3% 0.1 Mw - 0.1%

Table 4: The composition of some reference corpora.

Multilingual research projects aimed at comparing linguistic facts in different languages using corpora have to face to problem of comparability among the corpora. To obtain consistent results it is common practice to consider and use corpora with roughly the same subcorpus composition.

To favour the comparability issue in CORIS a further corpus - CODIS - was designed. Aimed at specialist needs arising in the context of interlinguistic analysis, CODIS presents a dynamic and adaptive structure that allows the selection of the subcorpora which are pertinent to a specific research project and also the size of every single sub-corpus. CODIS is designed to be dynamically adapted to different comparative needs. As shown in table 5, each CORIS subcorpus was split into four parts of different sizes. The sizes were carefully selected in order to allow, once combined in various ways, the creation of subcorpora of virtually any size. For example the subcorpus *Miscellanea* can be built of size 1, 2, 3 (2+1), 4, 5 (4+1), 6 (4+2), 7 (4+2+1), 8 (4+2+1+1) million words. This fine granularity create an extremely flexible corpus structure that can be adapted to almost any possible comparison with other reference corpora in different languages.

Subcorpus	**User-selectable sizes (Mw)**			
PRESS	16	8	4	2
FICTION	11	5	3	1
ACADEMIC PROSE	5	3	1	1
LEGAL & ADMIN. PROSE	4	2	1	1
MISCELLANEA	4	2	1	1
EPHEMERA	2	1	0.5	0.5

Table 5: CODIS user selectable subcorpora and their sizes.

5. Corpus access tools

In order to manage the huge amount of data involved in the creation of such large corpora, we need adequate computational procedures that have to be *general* – they have to accept different approaches to mark-up, tokenisation, languages, etc. – *flexible* – they must allow for corpus maintenance and adaptation – *user friendly*, and, last but not least, *extremely fast*. In response to these needs, O. Mason (1996) has devised CUE (Corpus Universal Examiner), a set of computer programs able to address all the requirements of a modern corpus retrieval application. The first version of CUE was written in C++ for UNIX systems, using the publicly available library Xforms (Zhao and Overmars 1995; Reichard and Johnson 1996) for the interface design. It involves fast procedures for the retrieval and access of data, and compression methods (Huffman coding) to reduce the amount of space needed to store the corpora. The main problem with this application was that it followed the stand-alone application paradigm. This meant that only the workstation that stored the corpora would have immediate access to them. Even if a complete Networked File System were provided, the application would run only on UNIX machines.

We transformed the stand-alone version of CUE into a client-server application in such a way that the server machine can provide corpus access across our Local Area Network. Moreover, we had to address a different problem, the multi-standard nature of our client workstation. At CILTA, we currently have Windows-based PCs, Macintoshes and UNIX workstations. It was not conceivable to develop and maintain a different client application for each kind of operating system/hardware platform pair. The natural, and unique, solution to such a problem was to develop the CUE client side in Java, obtaining, in theory, complete portability among different systems without any further effort.

The server side was derived from the original CUE release. It is written in C++ and runs on a Sun Enterprise Server 450 with 4 processors, 512MB of memory and 36GB of disk space supporting the Solaris 7 operating system. It was implemented following the concurrent server model so that it can accept multiple queries from different client machines at the same time. Once a new client makes a request to activate the service, a new copy of the server program is created and remains active until the client closes the connection. It is important to note that, for security reasons, the client has to provide authentication – as a legal JCUE client program – and the user, who is trying to access this service, has to provide passwords. In this way, we can restrict the use of some corpora to particular users or research teams.

The most complex work was to divide the stand-alone application into a server side and a client side, providing a complete set of operations needed to retrieve data from the network. We developed a scheme similar to Remote Procedure Call technique, building a client-and-server-module interface to the network communication protocol. These modules transform the request and the data from the client side into string codes that are sent across the network using the standard BSD socket support. Using a similar scheme, they transform the data retrieved by the server and send them back to the client.

The client side was completely redesigned using Java (version 1.2) and is currently working on Windows 95/NT PCs, Macintoshes, and Sun-Solaris UNIX workstations. We also developed an X-Window version of the client for UNIX machines, directly derived from the original CUE package.

Recent developments in corpus access tools underlined the needs for a web-based interface to corpus information retrieval. Using the same mechanism described above we produced a web-based-cgi client that connects the remote browser requests with the local server, querying the corpora and performing all the appropriate operations on concordances. The current version of the web-based client processes queries in one step, but we are devising a more complex and interactive application for better corpus access.

6. Part-of-speech tagging

Part-of-speech tagging and lemmatisation were immediately considered one of the main issues in the CORIS structure. As a consequence a separate project was started to produce a POS tagger for the Italian language. A lot of tagging programs were available freely on the Internet, but none of them was explicitly developed for the Italian language. Moreover, Italian is an highly inflected language, causing problems for taggers developed mainly for English language. This reason led us to devise a new tagger program designed especially for Italian language. The main difference from the most famous and most widely used taggers is the presence of a powerful morphological analyser based on a list of 100,000 lemmas, able to produce more then 1.7 million inflected forms. This huge lexicon enables us to cover almost the entire sum of words appeared in Italian texts, especially when compared with the finite lexicon produced exclusively from the training corpus. The words unknown to the tagger are thus reduced to proper nouns (78%), that are capitalised in Italian, common nouns (10%) and adjectives (7%). Starting from these proportions of unknown words and using a simple heuristic method we were able to solve about 95% of problems of this kind.

The theoretical model used for the tagger is the well known and widely used stochastic method based on hidden Markov models. Common techniques have been used to overcome the idiosyncrasies of HMM [Kempe 1993]: the influence of problems like numerical underflow in computations and sparse data matrix has been heavily reduced.

Two small corpora have been constructed to train the program and test the results of the CORIStagger and also to compare it with the most widely used retrainable tagging programs. The training corpus consists of about 84.000 words, while the test corpus is 22.000 words big; both corpora were manually tagged, to be sure about starting and control data.

The following table shows the classification error for each tagger considered in this study:

TAGGER	*ERROR*	*THEORETICAL METHOD*
CORISTagger	3.39%	Stochastic
TnT (Brants 2000)	4.39%	Stochastic
MXPOST (Ratnaparkhi 96)	5.27%	Maximum Entropy
Brill (Brill 92,94)	5.66%	Rule based
TreeTagger (Schmid 94)	9.24%	Stochastic

Table 6: Classification errors of Pos taggers considered in this study, when applied to unseen texts.

According with Brants (2000), "We have shown that a tagger based on Markov models yields state-of-the-art results, despite contrary claims found in the literature". Moreover, CORISTagger, being based on a large and powerful morphological analyser for lexicon management, performs best when tagging unseen texts.

Nevertheless, as outlined in Tamburini [2000], the tagset composition is one of the main issues affecting the tagger behaviour and tagging error. Currently we are studying statistical clustering methods for tagset composition, in order to choose the best compromise between linguistic needs and the reduction of automatic-tagging errors.

Note

The authors have elaborated every part of this paper together. As far as academic requirements are concerned R.Rossini Favretti takes official responsibility for sections 2.1, 2.2, 2.3, F. Tamburini for sections 4, 5, 6 and C. De Santis for section 3.

References

AIE 1999 *La produzione libraria italiana del 1997*. Milano.

Benjamin W 1966 *L'opera d'arte nell'epoca della sua riproducibilità tecnica*. Torino, Einaudi.

Bertola V 1999 *La rete e i diritti d'autore.* http://bertola.eu.org/icfaq/diritti.htm.

Biber D, Johansson S, Leech G, Conrad S, Finegan E 1999, *Longman Grammar of Spoken and Written English*. London, Longman.

Borges JL 1985 La biblioteca di Babele. In *Finzioni*. Torino, Einaudi.

Bortolini U, Tagliavini C, Zampolli A 1972 *Lessico di frequenza della lingua italiana contemporane.* Milano, IBM Italia.

Brants T 2000 TnT – A Statistical Part-of-Speech Tagger. In *Proc. Conference on Applied Natural Language Processin.* Seattle, WA.

Brill E 1992 A Simple rule-based part-of-speech tagger. In *Proc. of the Third Conference on Applied Natural Language Processing.* ACL, Trento.

Brill E 1994 Some advances in rule-based part of speech tagging. In *Proc. of the Third International Workshop on Parsing Technologies*. Tilburg, The Nederlands.

Calvo F, Ciotti F, Roncaglia G, Zela M 1999 *Internet 2000. Manuale per l'uso della rete*. Roma, Laterza.

Chimienti L 1999 *Lineamenti del nuovo diritto d'autore: direttive comunitarie e normativa interna*. Milano, Giuffrè.

De Mauro T, Mancini F, Vedovelli M, Voghera M 1993 *Lessico di frequenza dell'italiano parlato.* Milano, Etas Libri.

De Santis C 2000 Isole e tesori. Navigare alla ricerca di risorse per la costruzione di un corpus di italiano scritto. In Rossini Favretti R (ed) *Linguistica e informatica. Multimedialità, corpora e percorsi di apprendimento*. Bulzoni, Roma.

FIEG 1999 *La stampa in Italia 1995-1998.* Milano.

Francis WN, Kucera H 1964/1979, *Manual of information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Department of Linguistics, Brown University.

Kempe A 1993 A probabilistic tagger and an analysis of tagging errors. *Technical Report*. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

Johansson S, Leech GN, Goodluck H 1978, *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*. Department of English, University of Oslo.

Mason O 1996 Corpus access software: The CUE system. *TEXT Technology*, 6 (4): 257-266.

Marandola M 1996 *Diritto d'autore.* Roma, AIB.

Nencioni G 1983 Di scritto e di parlato. *Discorsi linguistici*. Bologna, Zanichelli.

Ratnaparkhi A 1996 A Maximum Entropy Model for Part-of-Speech Tagging. In *Proc. of Empirical Methods in Natural Language Processing Conference*. University of Pennsylvania.

Reichard K, Johnson EF 1996 Using Xforms. *Unix Review* 84.

Revelli C 1997 Discussioni sul copyright. *Biblioteche oggi* 10.

Rossini Favretti R 2000 Progettazione e costruzione di un corpus di italiano scritto: CORIS/CODIS. In Rossini Favretti R (ed) *Linguistica e informatica. Multimedialità, corpora e percorsi di apprendimento*. Bulzoni, Roma.

Rossini Favretti R 1998 Using multilingual parallel corpora for the analysis of legal language: the Bononia Legal Corpus. In Teubert W, Tognini Bonelli E, Volz N (eds), *Translation Equivalence. Proceedings of the Third European Seminar, Translation Equivalence.* The TELRI Association e.V., Institut für Deutsche Sprache, The Tuscan Word Centre, pp. 57-68.

Rossini Favretti R 1999 Equivalenze traduttive in corpora giuridici multilingue. *Quaderni di Libri e Riviste d'Italia 43. La traduzione IV*, Roma, Poligrafico e Zecca dello Stato, pp. 47-66.

Samuelson P 1991 Digital media and the law. *Communications of the ACM* 34 (10)

Schmid H 1994 Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proc. International Conference on New Methods in Language Processing*, Manchester, UK.

Sinclair J 1991 *Corpus, concordance, collocation*, Oxford, Oxford University Press.

Spina S 1997 *Parole in rete*, Firenze, La Nuova Italia.

Tamburini F 2000 Annotazione grammaticale e lemmatizzazione di corpora in italiano. In Rossini Favretti R (ed) *Linguistica e informatica. Multimedialità, corpora e percorsi di apprendimento*. Bulzoni, Roma.

Vitiello G 1997 Biblioteche, editoria e diritto d'autore, *Biblioteche Oggi* 1.

Zhao, TC, Overmars M (1995) *Forms Library. A graphical user interface toolkit for X.* http://bragg.phys.uwm.edu/xforms.