THE ITALIAN *PAROLE* CORPUS: AN OVERVIEW

RITA MARINELLI, LISA BIAGINI, REMO BINDI, SARA GOGGI, MONICA MONACHINI, PAOLA ORSOLINI, EUGENIO PICCHI, SERGIO ROSSI, NICOLETTA CALZOLARI, ANTONIO ZAMPOLLI

Abstract - *The PAROLE project (Preparatory Action for Linguistic Resources Organization for Language Engineering) has produced a set of harmonized corpora and lexicons for a large number of European languages. Each corpus, made up of 20 million words, was built up as reference corpus for Human Language Technology applications, to provide full information about a large variety of text types in the language considered, to represent the use of contemporary language and to become the first nucleus of an electronic text library. The texts have been stored using a common format following the standards recommended in the CES (Corpus Encoding Standard), according to flexibility and multifunctionality criteria. The texts belong to a wide range of media and genres, selected in proportions aimed at reflecting their prominence within the society, classified according to medium, genre, topic and time of production.*

Keywords - *textual resources, corpus design, corpus representation, corpus annotation*

1. INTRODUCTION

PAROLE was one of major projects launched by the EC for the construction of Language Resources (LR) in the field of written language. Over the last fifteen years there has been growing interest on the part of the NLP (Natural Language Processing) community towards the development of large reusable language data. The lack of big computational lexicons and the non-homogeneity of existing resources has been a hindrance to the progress of NLP applications. The LE-PAROLE project is aimed at building large, generic and reusable, uniformly structured textual and lexical databases for the European languages.

The project has been funded by EC DGXIII and implemented by the PAROLE Consortium coordinated by A. Zampolli. The group was initially formed by representatives of Institutes with recognized tradition of working with large textual corpora[1]. These Institutes began to cooperate by studying and comparing criteria for corpus composition, encoding formats, text annotation, etc., performing linguistic analyses to verify the possibility of harmonizing and coordinating their researches in the field of corpora.

After a series of workshops (Grosseto, 1987; Dubrovnik, 1988; Budapest, 1988), it was common view that top priority should be placed on the availability of adequate LRs. A feasibility study for setting up a "Network of European Reference Corpora" (NERC, promoted and coordinated by A. Zampolli, 1996; Calzolari *et al*., 1996) was proposed in Luxembourg by the Council of Europe Corpus Group. The major aims of the proposal were to provide textual corpora according to the needs of the European R&D (Research and Development) community and to agree on a common work-plan to guarantee the interoperability of the national reference corpora in the European multilingual context: "The need to ensure reusability, integration, global planning and coordinated international cooperation in the field of Language Resources has been stressed" (Calzolari and Zampolli, 1999).

2. PROJECT GENERAL LINES

In the work-plan of the PAROLE project stress has been put on producing written language resources suitable for different Language Engineering (LE) applications.

The lexicons are built around a generic model (an instantiation of EAGLES recommendations in an enriched GENELEX model), developing declarative, multifunctional lexical resources, able to easily evolve, for example by enlarging the lexical coverage or becoming multilingual.

[1]The Pisa group (A. Zampolli), the Institute Nationale de la Langue Française (InaLF, B. Quemada), the University of Birmingham (J. Sinclair), the Institut für Deutsche Sprache (W. Teubert), the University of Malaga (M. Alvar-Ezquerra).

The corpus part of the project - in which we are interested in this paper - was to produce large monolingual corpora of 20 million word tokens, in almost all of the 14 participating languages[2]. PAROLE[3] was probably the first project producing such large reference corpora in so many languages according to a common model.

The corpora are *harmonized*, i.e. built according to the same design and composition, following common EAGLES-PAROLE specifications and markup conventions.

The corpora have been assembled so as to be generic as well as usable for NLP training, testing and benchmarking purposes. The availability of large, uniformly structured resources "is a key condition for the feasibility of reaching the objectives established within the budget and time constraints" (Zampolli, 1997), offering the benefits of a standardized base allowing cheaper LE products supply.

The source texts composing the corpora are *encoded* following the standards recommended in the CES, produced by EAGLES on the basis of TEI (Text Encoding Initiative) guidelines.

A sub-corpus of 250,000 running words is *tagged* at the morphosyntactic level following the EAGLES guidelines.

Each partner of the project had the possibility to mark up and tag the corpus using its own software package, considering that the compatibility of the various corpora was ensured by the adoption of common criteria for composition, encoding and linguistic tagging.

[2] Italian, French, Greek, Catalan, English, Danish, Irish, Swedish, Finnish, Dutch, Belgian-French, Portuguese, Norwegian, German. In particular, 15 million words were produced for the Irish, 3 million words for the Norwegian corpus.

[3] PAROLE partners included the following countries: Italy, France, Greece, Spain, United Kingdom, Denmark, Ireland, Sweden, Finland, Holland, Belgium, Portugal, Norway, Germany.

3. GENERAL DESIGN CRITERIA

Harmonization, reusability and standardization have been the most important criteria in the design of the PAROLE corpora.

3.1. *Harmonization*

Harmonization means accordance with common criteria of composition and representation, following a specified data architecture and encoding formats. The corpora produced have been harmonized under the following aspects (LE-PAROLE, 1995: 18):

- overall corpus design and composition;
- data representation.

3.1.1. *Corpus composition*

As regards *composition*, the reference corpora must reflect contemporary language use, thus including basic reference material. A reference corpus[4] can be the base for a very wide range of applications, concerning contemporary languages. It was therefore necessary to collect texts of different types, "a collection of a broad variety of written material reflecting language variety" (ibid.: 19).

As to composition, the corpus texts were selected and grouped so that the corpora would share the same features, namely:

a) they include texts belonging to a wide range of media and genres, selected according to proportions defined in the project for all languages and reflecting their prominence in society (PAROLE MLAP, 1996);
b) they incorporate texts of a broad range of common topics, with

[4] A reference corpus is a "general purpose corpus that has acquired a certain definitive status with respect to a particular language at a particular time in its history" ; or "a corpus designed to provide comprehensive information about a language" (LE PAROLE, 1995: 18).

examples from all major topic areas, in order to support the building of generic lexicons;

c) they do not include texts written by specialists for specialists in the same field;
d) they do not include texts older than 1970;
e) they do not include transcribed spontaneous speech.

Constructed on similar principles, the PAROLE reference corpora in several languages form a group of comparable corpora[5], because of the coordination of medium, genre, topic, size and - to some extent - year of production[6].

3.2. *Reusability*

In the last few years special attention was paid to the notion of *reusability* both in the sense of reusing existing language resources (machine-readable dictionaries, text corpora), and in the sense of building language resources suitable to be used in many different theoretical and application frameworks. Great incentive was put on producing generic Written Language Resources (WLR), suitable for LE. Such an effort was also based on the reuse of existing partial language resources, prepared and converted to become reference material.

3.3. *Standard data representation and text classification*

The development of a 'common encoding system' (for *interoperability* and data exchange) is closely linked to the concept of *harmonization* and *reusability*. The PAROLE corpora are harmonized also with respect to text representation: this means that the data are encoded using a common format, based on CES/TEI (Text Encoding Initiative) and on EAGLES/Multext conventions: "The PAROLE standard follows the recommendations made by

[5] "A comparable corpus is a corpus which selects similar texts in more than one language or variety" (Sinclair, 1994).
[6] See NERC recommendations (Calzolari *et al.*, 1995) in which the main selection principles are "topic or subject matter" and "text medium".

EAGLES... a text encoded according to the PAROLE standard will parse with the TEI DTD (Document Type Definition)" (Ridings, 1996).

The CES "specifies a minimal encoding level that corpora must achieve to be considered standardized", in terms of descriptive representation as well as general architecture (Corpus Encoding Standard, 1996: 1)[7]. A PAROLE DTD was designed and utilized by all the partners. The use of a common DTD ensures the formal coherence in encoding the various texts within a corpus and between the various corpora.

A uniform system of features was designed and agreed. The TEI recommendations state that for large corpora such uniform annotation is feasible only restricting the number of distinctive features to those which are linguistically more relevant.

Data representation must however be flexible and multifunctional: data description choices must be variably grained and fully documented. A classification at two levels was decided: a general level that is mandatory and an optional detailed level. A few mandatory formalised descriptors are at the basis of laying down the harmonized composition of the reference corpus: Medium, Genre, Topic, and Time of production. They are found in the *header*, an ordered collection of attributes, each with a set of values, at the top of each PAROLE text, as specified below (section 5.2.2).

PAROLE has agreed on a set of parameters to be used in the headers. With respect to the parameter medium, which records the mode of transmission for which the text was originally composed (Sinclair and Ball, 1995: 8-9), each text is classified according to the following categories of medium:

- Book
- Newspaper
- Periodical

[7] The CES is an application of SGML (ISO 8879: 1986, Information Processing-Text and Office Systems-Standard Generalized Markup Language) compliant with the specifications of TEI Guidelines for "Electronic Text Encoding and Interchange".

- Miscellaneous (correspondence, electronic, ephemera, hand-written, typed, others).

The *medium* is the key-parameter to defining the common overall quantitative composition of the corpora. It was agreed that the overall composition of the corpora was to comply with the following distribution according to the media:

Medium	Percentage	
	Min	Max
Book	16%	22%
Newspaper	58%	72%
Periodical	4%	10%
Miscellaneous	8%	12%
TOTAL	100%	

Figure 1

4. AVAILABILITY OF THE CORPORA

For each of the participating languages it was agreed in the Consortium that the overall Corpus, of at least 20 million words, was to be available at the partner's site; the so-called Distribution Sub-Corpus, of at least 3 million words, was to be available for distribution to the outside community; the Linguistically Annotated Sub-Corpus, of at least 250,000 words, annotated at the morphosyntactic level, was also available for distribution.

5. THE ITALIAN PAROLE CORPUS

We give in this section a few details on the Italian PAROLE Corpus.

5.1. *Composition*

The composition of the Italian PAROLE corpus is reported below:

Medium	Number of words	Percentage of total corpus
BOOKS	3,752,643	17.91%
NEWSPAPERS	14,596,649	69.68%
PERIODICALS	959,255	4.58%
MISCELLANEA	1,640,189	7.83%
TOTAL	20,948,736	100%

Figure 2

The distribution of the material within each *medium* is shown in the next sections, mentioning the sources and the state of legal agreement with copyright holders, a very important aspect for data reusability (the quantitative data appear in detail in the Appendix).

All the data were collected from versions of the texts recorded on magnetic supports of various type. The texts collected covered a period of more than 25 years: from 1970 to 1997 (the diachronic distribution of the texts can be seen in the Appendix).

5.1.1. *Books*

Also with regard to the books, we exploited the electronic sources already used to build up the ILC Italian Reference Corpus. Only four new books were added to the existing data to reach the required amount of words; 44 entire books were processed obtaining a total of 3,752,643 words.

Books

Source	Time span/ Date	State of legal agreement with copyright holders	Current number of words
Mondadori	1970-1995	Signed, with restrictions (for research only)	619,651
Einaudi	1970-1995	Signed, with restrictions	3,132,992
Subtotal			3,752,643

Figure 3

5.1.2. *Newspapers*

As regards the Newspapers, the choice of the issues aimed at ensuring a regular distribution of the material over a 5-year period, covering all the days of the week. The issues deal regularly with different topics in different weekdays, therefore special attention was paid to prevent issues from different newspapers chosen from the same day.

Entire newspaper issues and articles were included. For each newspaper, the number of words per year is equally distributed over a 12-month range.

Newspapers

Source	Time span/ Date	State of legal agreement with copyright holders	Current number of words
La Repubblica	1992-1996	Signed, with restrictions (for research only)	3,383,435
Il Corriere della Sera	1992-1996	Signed, with restrictions	3,232,598
La Stampa	1992-1996	Signed, with restrictions	3,293,704
Il Sole-24 Ore	1992-1996	Signed, with restrictions	4,153,131
L'Unione Sarda	March 1996	Signed, with restrictions	533,781
Sub-total			14,596,649

Figure 4

5.1.3. *Periodicals*

As regards the Periodicals, 9 of these were selected which covered both general and particular topics.

Two issues per year for each periodical were chosen, covering a 4-year time period (1985-1988). All these texts were selected from the data of the pre-existing ILC Italian Reference Corpus (Bindi *et al.*, 1989) applying the *reusability* criterion.

The total amount of words (more than 950,000) is distributed over the nine periodicals (ca. 100,000 words each).

Periodicals

Source	Time span/ Date	State of legal agreement with copyright holders	Current number of words
Mondadori Editore Milano	1985-1988	Signed (restricted for research only)	959,255
Sub-total			959,255

Figure 5

The average length of the newspaper and periodical articles appears at the end of the Appendix.

5.1.4. *Miscellaneous*

The Miscellaneous material of the Italian Corpus contains over 1,600,000 words, corresponding to 8% of the entire Corpus. All the data were collected according to the categories of the classification feature Miscellaneous (correspondence, electronics, ephemera, hand-written and typed material), and classified according to the year of production, genre, topic, following the obligatory PAROLE features for text classification (PAROLE MLAP, 1996: 6). This was not always easy owing to the particular nature of some documents.

The sources of the documents were various and of different types, as seen in figure 6.

Miscellaneous

Source	Time span/ Date	State of legal agreement with copyright holders	Current number of words
CNR:			
-Rules & decrees	1992-1993		453,843
-Research projects	1992	Full	887,103
-Patents	1987-1991		60,276
-Research activity	1995		89,303
Livorno Town Council:			
- Press releases	1996		55,505
- Communications,		Full	
instruction notes, etc.	1997		39,960
Teatro Verdi of Pisa:			
- Legal documents, minutes,		Full	
reports, notes	1995-1997		21,219
ASAMAR (Association of Livorno shipping Agents):			
- Letters,circulars,press releases	1996-97	Full	32,980
Sub-total			1,640,189

Figure 6

Great effort was made to unify all the data formats into a common "text" format (ASCII), which became the basis for the following procedures.

5.2. *Text representation*

5.2.1. *Intermediate text encoding in DBT format*

The data of all Italian texts of the PAROLE corpus were converted in SGML format after an intermediate conversion phase in DBT (Textual Data Base) format. DBT (see Picchi, this volume) is a system created at ILC to manage large Language Resources. After the automatic conversion of the source in the format required by the DBT system, the texts were checked manually. A spelling checker was used to identify and correct spelling errors in the source. In this intermediate format (DBT) two types of information were already supplied: the *header*, with information separated from the text itself, but useful for its (bibliographic) classification; and the *body* of the text, with information about the segmentation of the

text itself and about the properties and characteristics of its segments (section, paragraph, sentence, word). For these two main groups of information (header and body of the text), it was decided to encode both the obligatory and some of the facultative features, in order to produce a fine-grained annotated corpus.

A detailed list of all the DBT and SGML tags used for each *medium* is given in the LE-PAROLE Italian Corpus Documentation (Goggi *et al.*,1997).

5.2.2. *SGML representation of the corpus*

SGML (Standard Generalized Mark-up Language) is the language to represent the standard encoding recommendation for textual corpora (CES). The TEI Guidelines use SGML to define a set of comprehensive conventions for representing documents.

SGML was adopted as an interchange format within the PAROLE project; therefore a DBT towards CES/SGML conversion procedure was set up and used to convert DBT format into SGML format.

All the standard elements conformant with the PAROLE DTD were stored in the *header* of each text, codified following the specified rules, providing administrative information (editorial principles, responsible for the corpus collection, etc.), bibliographical information allowing the identification of the text (name of authors, title, year of publication, etc.) and information on the text type, which classifies the text according to a system of descriptive parameters (e.g. medium, genre, topic, etc.).

The corpora are usually formed by relative proportions of one or more text types: the descriptive parameters contained in the *header* allow the selection of sub-corpora for specific tasks, specific domains, etc.

Additional criteria can be found in the *header*: in fact each partner could adopt additional classifications features, based on specific requirements: scientific criteria, cultural situation, etc. With regard to the Italian texts, the lists containing all the types of keywords, giving extra information, encoded for newspapers, periodicals and miscellaneous can be found in the LE-PAROLE Italian Corpus Documentation (ibid.: 1997).

5.3. *Linguistic annotation of the texts*

The subset of 250,000 word-forms of the PAROLE corpus to be annotated linguistically was extracted from a set of newspapers and periodicals. Linguistic annotation was performed in various steps.

A first annotation was performed automatically by Pi-Morpho, the morphological engine of the Pi-System, based on the Italian lexical component (DMI), that assigns to each word-form in the text all possible morphosyntactic interpretations augmented with *lemma* information. A further module, the Pi-Tagger (Picchi, 1994), was used to compute the most likely interpretation among all possible alternatives, with a success rate of 97%.

The linguistic specifications for tagging were based on EAGLES recommendations for morphosyntactic annotation (Monachini and Calzolari, 1996). The set of tags applied is the one used at ILC, mappable on the tag set agreed on within the PAROLE Project (Corazzari *et al*., 1996).

Automatic annotation was followed by an appropriate interactive post-tagging procedure, Tagg-Hand, aimed at checking and correcting automatic annotation. A level of maximum granularity was reached in annotation: information was checked up to the level of morphological features (cf. a list of information types encoded in the next page).

An *ad-hoc* automatic procedure was designed to improve the correctness and consistency of tagging. This procedure highlighted 'non-admitted' tags within the chosen tag set, signalling potential mistakes that the annotators could correct. Furthermore, it extracted lists of triplets 'lemma-word form-tag': these lists were of extreme usefulness since, once sorted either by lemma, word form or tag, they made it possible to perform consistency checking in annotation. To give an example, by sorting the triplets in alphabetical order of the word form, it could be observed whether the same word forms received the same interpretation; conversely, by sorting on the tag, it could be checked whether a tag was consistently applied to the same set of linguistic phenomena.

Finally, when the tagging was considered correct, the procedure, carried out the conversion of the ILC tag set into the final PAROLE tag set. This conversion routine was further extended

to cover the tag sets used within the framework of other projects, i.e. MULTEXT and ELSNET, allowing their conversion into the PAROLE tag set. The aim was to disseminate the PAROLE results and to produce shareable textual resources annotated with harmonized tag sets to be put at the scientific community's disposal.

PoS	**Part-of-Speech**
adj	adjective
adv	adverb
conj	conjunction
det	determiner
adp	prepostion
intj	interjection
num	numeral
pron	pronoun
art	article
noun	noun
resd	residual
abbr	abbreviation
verb	verb
degree	**degree**
pos	positive
sup	superlative
type	**type**
frgn	foreign
coord	coordinative
subord	subordinative
dem	demonstrative
escl	esclamative
indf	indefinite
poss	possessive
rel	relative
int	interrogative
prep	preposition
uspc	underspecified
card	cardinal
ord	ordinal
pers	personal
def	definite
comm	common
prop	proper
gend	**gender**
fem	feminine
masc	masculine
comm	common
numb	**number**
sing	singular
plur	plural
inv	invariant
inloc	**in locution**
y	found in loc.
n	not-found in l.
form	**formation**
simple	simple
comp	coumpound
pers	**person**
1	1st
2	2nd
3	3rd
clit	**presence of clitics**
y	verb with clitic
n	verb without clitic
vf-m	**verbform/mood**
ger	gerund
inf	infinitive
ind	indicative
subj	subjunctive
cond	conditional
impr	imperative
part	participle
tense	**tense**
pres	present
impf	imperfect
fut	future
past	past

Figure 7

5.4. *Enriching the PAROLE corpus*

The Italian PAROLE Corpus consists of 20 million word tokens, with texts collected until 1997. After the end of the PAROLE Project, the Corpus has been enlarged up to 73 million words, adding data from 4 newspapers, covering 5 years from 1997 to 2001. All data were encoded following the general standard rules recommended by the Project.

The table here below represents the part of the data that has been added.

Year	N. Words
1992	4,293,422
1993	4,165,250
1994	4,163,950
1995	4,105,030
1996	4,642,189
1997	7,089,055
1998	7,087,026
1999	7,075,689
2000	7,526,907
2001	2,063,168
TOTAL	52,211,686

Figure 8

The agreement for the use is not full: at present the material can be treated only for internal use.

6. CONCLUSIONS

We have first described the general lines of the PAROLE Corpus project and the Italian Corpus in detail. One of the main goals of the LE-PAROLE project was to ensure the creation of a comparable set of large WLRs for all the EU languages.

The purpose of the project was not to create new technology, but to produce and make the textual corpora available to the

scientific community in a standard form, leading to reuse and interoperability.

But enterprises like this one never end. After the conclusion of the Project, the Corpus size will be more than three times the original PAROLE size, still conforming the design criteria described before. Further enrichment is expected in the future. The results of LE-PAROLE, enlarged and maintained, offer a significant contribution to the industrialization of the LE-sector. These resources built within LE-PAROLE Project constitute a important cooperative step towards the harmonization of language policies across the European Union.

APPENDIX

Quantitative Data

Newspapers

	Current n. of Words
La Stampa	3,293,704
La Repubblica	3,383,435
Il Corriere della Sera	3,232,598
Unione Sarda	533,781
Il Sole-24 Ore	4,153,131
Total for Newspapers	14,596,649 (69,68%)

Periodicals

	Current n. of Words
Casaviva 1985-88	114,849
Cento cose 1985-88	116,473
Epoca 1985-88	120,405
Espansione 1985-88	78,581
Grazia 1985-88	111,196
Panorama 1985-88	82,403
Starbene 1985-88	119,799
Storia illustrata 1985-88	114,154
Zerouno 1985-88	101,395
Total for Periodicals	959,255 (4.58%)

Miscellaneous

	Current n. of Words
CNR Rules/Decrees	453,843
CNR Projects 1992	887,103
CNR Projects 1995	89,303
CNR Patents 1987-91	60,276
Total for CNR	1,490,525
Pisa - Teatro Verdi	21,219
Livorno - ASAMAR	32,980
Livorno - Town Council Various Material	39,960
Livorno - Town Council Press Releases	55,505
Total	149,664
Total for Miscellaneous	1,640,189 (7.83%)

Books

	Current n. of Words
Mondadori	619,651
Einaudi	3,132,992
Total for Books	3,752,643 (17.91%)

Total amount of words 20,948,736

Diachronic Distribution of the Texts

Year	News-papers	Periodi-cals	Books	Miscella-Neous	Total	%	Decade
1970			36,440		36,440	0,17	
1974			122,416		122,416	0,58	70-79
1977			304,616		304,616	1,45	2,20%
1980			58,108		58,108	0,28	
1982			76,900		76,900	0,37	
1985		284,189	599,051		883,240	4,22	
1986		240,092	683,064		923,156	4,41	80-89
1987		202,587	458,892	13,436	674,915	3,22	22,44%
1988		232,387	1,111,783		1,344,170	6,42	
1989			301,373	16,793	318,166	1,52	
1990				14,504	14,504	0,07	
1991				15,543	15,543	0,07	
1992	2,757,213			1,036,740	3,793,953	18,11	
1993	2,875,370			304,206	3,179,576	15,18	
1994	2,994,561				2,994,561	14,29	90-97
1995	2,962,397			89,303	3,051,700	14,57	75,36%
1996	3,007,108			78,115	3,085,223	14,73	
1997				71,549	71,549	0,34	
Tot.	14,596,649	959,255	3,752,643	1,640,189	20,948,736		100%
%	69,68	4,58	17,91	7,83		100	

Average Length of a Newspaper Article (in words)

Newspaper	N. of Words	N. of Articles	Average n. of Words in an Article
Il Corriere della Sera	3,232,598	7,654	422
Il Sole-24 Ore	4,153,131	7,643	543
L'Unione Sarda	533,781	1,522	351
La Repubblica	3,383,435	6,034	561
La Stampa	3,293,704	7,546	437
Total	14,596,649	30,399	480

Average Length of a Periodical Article (in words)

Periodical	N. of words	N. of articles	Average n. of Words in an Article
Casaviva	114,849	69	1,664
Cento cose	116,473	89	1,309
Epoca	120,405	84	1,433
Espansione	78,581	43	1,827
Grazia	111,196	91	1,222
Panorama	82,403	90	916
Starbene	119,799	77	1,556
Storia Illustrata	114,154	50	2,283
Zerouno	101,395	61	1,662
Total	959,255	654	1,467

REFERENCES

ATKINS S., CLEAR J., OSTLER N., *Corpus design criteria,* «Literary and Linguistic Computing», VII (1992), 1, 1-16.

BINDI R., MONACHINI M., ORSOLINI P., *Il Corpus di Riferimento della lingua italiana contemporanea, Documentazione,* ILC-TLN-1, 1989.

BURNARD L., *Text Encoding for Information Interchange - An Introduction to the Text Encoding Initiative,* Oxford University Computing Services, 1995.

CALZOLARI N., BAKER M., KRUYT T. (eds.), *Towards a Network of European Reference Corpora: Report of the NERC Consortium Feasibility Study,* Coordinated by A. Zampolli, «Linguistica Computazionale», XI (1995).

CORAZZARI O., MONACHINI M., ROVENTINI A., *Italian Morphosyntactic Tagset: Guidelines for the Interpretation and the Manual Checking,* PAROLE Deliverable, Pisa, 1996.

CORPUS ENCODING STANDARD, *Document CES* 1, 1996.

GENELEX CONSORTIUM, *EUREKA Project GENELEX - Report of Syntactic Layer,* 4.0., 1993.

ISO 8879, *Information Processing - Text and Office Systems - Standard Generalized Mark Up Language (SGML),* ISO, Geneva, 1986.

ISO/IEC 10646-1, *Information Technology - Character Sets and Information Coding,* 1986.

GOGGI S., BIAGINI L., BINDI R., MARINELLI R., PICCHI E., ROSSI S., *LE-PAROLE Italian Corpus Documentation,* P-WP2.11-MEMO-PSA-3, 1997.

LE-PAROLE, *Technical and Financial Annex,* LE-4017, 1995.

LE-PAROLE, *Addendum to Technical Annex,* WP 4, 1996.

MONACHINI M., CALZOLARI N., *Synopsis and Comparison of Morphosyntactic Phoenomena encoded in Lexicons and Corpora. A common Proposal and Application to EU Languages,* EAGLES Recommendations, Pisa, ILC, 1996.

PAROLE MLAP. 63-386, *Design and Composition of Reusable Harmonized Written Language Reference Corpora for European Languages,* prepared by Ole Norling-Christensen, 1996.

PICCHI E., *D.B.T.: A Textual Data Base System,* in Computational Lexicology and Lexicography, Special Issue dedicated to B. Quemada, *Linguistica Computazionale*, VII (1991), 177-205.

PICCHI E., *Statistical Tools for Corpus Analysis: a Tagger and Lemmatizer for Italian, Proceedings of EURALEX'94*, Amsterdam, 1994.

RIDINGS D., *Text Representation in PAROLE*, Goteborg, 1996.
SINCLAIR J., *Corpus Typology*, EAGLES Document EAG-CSG/IR-T1.1, 1994.
SINCLAIR J., BALL J., *EAGLES Text Typology*, 1995, 8-9.
WALKER D., ZAMPOLLI A., CALZOLARI N. (eds.), *Automating the Lexicon: Research and Practice in a Multilingual Environment, Proceedings of a Workshop held in Grosseto*, Oxford, 1995.
ZAMPOLLI A., HOVY E., *Government: Policies and Funding*, in E. HOVY, N. IDE, R. FREDERKING, J. MARIANI, A. ZAMPOLLI (eds.), *Multilingual Information Management: Current Levels and Future Abilities, a Report Commissioned by the US National Science Foundation*, Vassar College, 1999, 129-136.
ZAMPOLLI A., *Introduction* in N. CALZOLARI, M. BAKER, T. KRUYT, (eds.), *Towards a Network of European Reference Corpora: Report of the NERC Consortium Feasibility Study*, Coordinated by A. Zampolli, «Linguistica Computazionale», XI (1995).
ZAMPOLLI A., *The PAROLE Project* in R. MARCINKEVICIENE, N. VOLZ (eds.), *The General Context of the European Actions for Language Resources, Second European Seminar: Language Applications for Multilingual Europe*, TELRI, Kaunas, Lithuania, 1997, 185-210.