Introducing the *La Repubblica* Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian

**Marco Baroni, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni,
Alessandra Volpi, Guy Aston, Marco Mazzoleni**

SSLMIT, University of Bologna
Corso della Repubblica 136, 47100 Forlì, Italy
{baroni, silvia, fcomastri, lpiccio, avolpi, guy, mazzoleni}@sslmit.unibo.it

Abstract

This paper describes the *La Repubblica* corpus, currently being developed at the SSLMIT of the University of Bologna. The corpus is a very large collection of newspaper text, currently amounting to 175 million words, but expected to grow to 400 million before the end of 2004. When completed, it will contain all the articles published between 1985 and 2000 by the national daily *La Repubblica*. The paper discusses the techniques used to extract the text, tokenize it and annotate it (basic TEI annotation, POS tagging, genre/topic categorization), it presents examples of how it can be used, and gives details of the ways in which interested users can access it. The paper concludes with a discussion of current and future developments, and of weak and strong points of this resource.

1. Introduction

This paper describes the *La Repubblica* corpus, currently being developed at the SSLMIT of the University of Bologna.

The corpus is a very large collection of newspaper text, currently amounting to about 175 million words, but expected to grow to 400 million words by the end of this year. When completed, it will contain all the articles published between 1985 and 2000 by the national daily *La Repubblica*, the second most widely-read Italian newspaper.

The texts in the corpus are POS-tagged and categorized in terms of genre and topic. All the information is encoded in XML following the TEI standards.[1]

This resource answers a widely-felt need for annotated contemporary Italian language data. While arguably not ideal as a reference corpus – being mono-source – the *La Repubblica* corpus is probably the largest freely accessible Italian corpus available to date (see Biagini et al. (2000) and Rossini Favretti et al. (2002) for sizeable alternatives).

In this paper we first provide an overview of the corpus and we describe the techniques used to construct it and to annotate it. We then discuss its availability and the current and planned access options. Next, we provide several examples of searches users can perform on the corpus. Lastly, we discuss some strong and weak points of the corpus, and look at improvements/enlargements currently being implemented and planned for the future.

2. Overview

Table 1 summarizes the characteristics of the *La Repubblica* corpus in its current stage (late February 2004) and in the stage we envisage for the end of 2004.

	current status	*end of 2004*
years	8	16
issues	2,085	5,163
articles	224,140	593,593
tokens	175,239,348	∼400M
types	880,111	?
sentences	6,316,532	?
pos tagging	yes	yes
lemmatization	no	yes
categorization	yes	yes
sara client access	partial	yes
cqp web interface	yes	yes
cwb-scan-corpus web interface	yes	yes
web service	no	?

Table 1: *La Repubblica* corpus overview

3. Corpus construction and annotation

3.1. Data extraction, tokenization and structural annotation

The texts that form the corpus were originally published as a series of CD-ROMs with their own search software. The articles and any available pieces of information (byline, title, subtitle, date and page number in the source newspaper) were extracted from a binary database to ASCII files, one file per newspaper issue. Meta-textual data not directly available from the database (i.e., those accessible only through the CD software) were discarded.

The resulting texts were tokenized and normalized in a number of ways, using regular expressions and manual checking.

Next, sentence boundaries were identified with manually crafted rules. Since such rules proved inadequate for segmenting titles, the latter have not been analyzed into sentences nor processed further, for the time being.

Since no information about paragraphing could be gleaned from the source texts, each article was treated as one paragraph.

As a first step in corpus annotation, the meta-textual data extracted from the CDs and the basic structural information identified automatically were added to the corpus following the TEI guidelines for text encoding and interchange. The resulting corpus has a header, which contains information about the collection as a whole (i.e., contents, authors, copyright restrictions, editorial interventions, updates etc.), and a series of texts, each corresponding to one newspaper issue and preceded by its own minimal header containing the publication date for that issue and basic copyright information. The text is further subdivided into

[1] `http://www.tei-c.org/P4X`

numbered *div* elements, corresponding to single articles. Each *div* is structured as follows: a title, a non-obligatory subtitle, a by-line, and a series of numbered *s* units, i.e., the automatically-identified sentences.

The second step in corpus annotation involved adding part-of-speech (POS) tagging and categorizing each article in terms of its genre and topic. Both tasks were carried out using supervised machine learning techniques. In this respect, annotation of a very large corpus also proved to be an ideal testbed for recent tagging and categorization algorithms.

3.2. POS tagging

3.2.1. Tagset

We originally designed a tagset that was in accordance with the EAGLES guidelines (Monachini, 1996). However, in preliminary experiments we realized that some of the distinctions made by EAGLES were not supported by distributional evidence. Thus, they were seriously harming the performance of our taggers.

For this reason, we developed the experimental 50-category tagset presented in table 2.

tag	*description*	*tag*	*description*
ADJ	adjective	ADJ:abr	adjectival abbreviation
ADV	adverb	ADV:abr	adverbial abbreviation
ART	article	ASP:fin	aspect. verb fin. form
AUX:fin	aux. verb fin. form	AUX:geru	aux. verb gerundive
AUX:infi	aux. verb infinitive	AUX:pper	aux. verb past part.
CAU:fin	caus. verb fin. form	CAU:geru	caus. verb gerund.
CAU:infi	caus. verb infinitive	CAU:pper	caus. verb past part.
CLI:ne	ne clitic	CLI:si	si clitic
CON:coo	coordinating conj.	CON:sub	subordinating conj.
DET:demo	demonstrative det.	DET:indef	indefinite determiner
DET:num	numeral determiner	DET:poss	possessive determiner
DET:wh	wh determiner	INT	interjection
LOA	loan word	MOD:fin	modal verb fin. form
MOD:infi	mod. verb infinitive	MOD:pper	mod. verb past part.
NOM	noun	NOM:abr	nominal abbreviation
NPR	proper noun	NPR:abr	proper noun abbrev.
NUM	number	PON	punctuation mark
PRE	preposition	PRE:art	prep. with article
PRO:demo	demonstrative pron.	PRO:indef	indefinite pronoun
PRO:num	numeral pronoun	PRO:pers	personal pronoun
PRO:poss	possessive pronoun	PRO:wh	wh pronoun
SENT	sentence marker	UNK	unknown
VER:fin	verb finite form	VER:geru	verb gerundive
VER:infi	verb infinitival	VER:pper	past participle
VER:ppre	present participle	WH	wh element

Table 2: Tagset

Among the non-EAGLES-conformant choices we made, we merged interrogative and relative pronouns (and determiners) into a single class: PRO:wh (DET:wh). Also, we decided to group all the instances of words such as *dove* "where", *come* "how" and *perché* "why" into the WH class.

Notice that our categories tend to be closer to those postulated in modern work on syntax (e.g., Graffi (1994)) than the more traditional EAGLES categories.

Tamburini (2000) reports improvements in Italian tagging performance with a tagset which is even more radically based on distributional (as opposed to morphosemantic) grounds than ours. We plan to experiment with a tagset along such lines in future research.

3.2.2. Training, testing and tagging

We first tagged a set of 180 randomly selected articles (about 115,000 tokens) using the pre-trained Italian version of the TreeTagger (Schmid, 1994). The output was converted into our tagset and revised by hand.

This manually cleaned corpus was used to experiment with single taggers and combinations of taggers from the ACOPOST (formerly ICOPOST) suite (Schroeder, 2002).[2]

Table 3 reports statistics about the percentage word-level accuracy achieved by the three ACOPOST taggers we used and by the best performing combinations in a series of 10-fold cross-validation tests.

	HMM	TBT	ET	COMB	STCOMB
Min	93.27	93.74	91.82	94.13	94.12
Median	95.04	95.12	94.00	95.63	95.71
Mean	94.81	94.96	93.72	95.44	95.46
Max	95.85	96.08	95.22	96.46	96.36

Table 3: Performance of taggers

HMM is a Markov Model tagger, TBT is a transformation-based tagger and ET is an example-based tagger.

COMB is a majority voter that, in case of ties, picks the HMM tag. STCOMB (for STacked COMBination) is a majority voter that uses a HMM-TBT stack instead of TBT (i.e., the TBT tagger takes a corpus pre-tagged by the HMM tagger as input). STCOMB also picks the HMM tag in case of ties.

In general, tagger combinations performed better than single taggers. The best performance was achieved by STCOMB, with a mean word-level accuracy of 95.46% in the 10-fold experiments. As far as we know, this is around the state-of-the-art performance level for tagging of Italian (Tamburini (2000) reports an accuracy of 96.61% with a much smaller, 21-category tagset).

We used STCOMB, trained on the full manually annotated set, to tag the remainder of the corpus.

3.3. Genre and topic categorization

Categorization in terms of genre and topic was performed using support vector machines as implemented in the SVMLight Package (Joachims, 1999).

We chose support vector machines since they are consistently reported to be among the best performing algorithms in text categorization (Sebastiani, 2002), and because they do not require preliminary selection of the features to be employed in classification. This is advantageous both in terms of practical development (the data can be directly fed into the algorithm without preprocessing) and, more importantly, in terms of performance maximization, since we can be confident that we are not discarding potentially relevant data (Joachims, 1997).

We created a manually annotated training set of 15,000 articles that were categorized into two genres (*news-report* vs. *comment*) and ten topics (*church, culture, economics, education, news, politics, science, society, sport, weather*).

10-fold cross validation tests on this set indicated that a simple non-lemmatized unigram TFIDF[3] model based on all the words occurring in an article (including author, year

[2] On tagger combinations, see, for example, van Halteren et al. (2001).

[3] Term frequency times inverted document frequency.

and title) performed rather well in both genre and topic detection.

In the genre detection experiments, this approach achieved an average accuracy of 90.03% with 90.89% precision and 93.03% recall. In topic detection, it achieved an average accuracy of 95.75% with 86.05% precision and 73.4% recall (measures micro-averaged across categories).

We also experimented with combined genre-topic categories, i.e., we trained the algorithm with categories such as *news-report/economics* and *comment/economics*. However, the results we obtained were worse than when treating genre and topic detection as independent tasks (96.54% accuracy,[4] 79.08% precision, 50.09% recall).

The genre detection results are of particular interest, since the performance we attained is comparable to the one reported in genre detection experiments with more sophisticated feature sets that include part of speech information, text statistics and/or ngrams: See, e.g., Pang et al. (2002) and Finn and Kushmerick (2003).

We suspect that the success of our simple approach based on non-lemmatized unigrams depends on the fact that Italian is an inflectionally rich language. Thus, non-lemmatized unigrams carry explicit morphological cues of different genres (e.g., first person and conditional verb inflections signal subjective styles). In this perspective, we expect that if we repeat the experiments with lemmatized unigrams, topic detection performance will improve (since topic detection depends on lexical cues) but genre detection performance will drop. We plan to test this hypothesis in further studies.

The model trained on the entire manually tagged set was used to categorize the remainder of the corpus.

4. Availability and access options

Under the terms of our agreement with La Repubblica, the full corpus cannot be distributed, and access to most information is only allowed for non-commercial purposes. For details on how to obtain access to the corpus, and for up-to-date information on annotation status and access options, please visit the following site:

`http://sslmit.unibo.it/repubblica`

Corpus data are available via through several interfaces.

They can be accessed from Windows machines through the free SARA client.[5] The version of the corpus that can be accessed via SARA at the moment lacks POS and genre/topic information.

The full corpus (including POS and genre/topic information) can also be accessed online with a web browser, in a version that has been indexed with the IMS Corpus WorkBench (CWB) and that can be queried using the Corpus Query Processor (CQP) language (Christ, 1994).

These interfaces are useful for a qualitative analysis of patterns in the corpus and to extract basic quantitative information. However, they are probably not of much use to those who are interested in extracting large-scale quantitative data (e.g., various forms of frequency lists to be used, say, in a machine learning task).

For these purposes, we provide a web interface to cwb-scan-corpus, where the output of a query is automatically sent to the issuer (cwb-scan-corpus is a CWB tool that produces frequency lists for ngrams matching a certain pattern).

We are also studying ways to offer more flexible public access to distributional information while respecting our agreement with the newspaper company. In the longer term, this could take the form of a web service API, so that interested users can access corpus information directly from their programs.

Lastly, pre-compiled unigram frequency lists extracted from the corpus and ACOPOST models trained on our manually annotated data are available for download (see the URL given above for updates).

5. Examples

Besides simple word or phrase searches on the whole corpus, the annotation and software provided allow the user, among other things, to limit searches to specific parts of the corpus (e.g., titles vs. bodies of articles, beginning vs. end of sentences etc.), to select sub-corpora on the basis of information contained within certain tags (e.g., all articles by a given author, all articles published in December, all sport news-reports etc.), to sort solutions according to the part of speech of the search word and/or of words in its context and so forth. Tables 4, 5, and 6 give a glimpse of this potential.

Table 4 compares the words tagged as adjectives which occur immediately to the left and to the right of the word forms *opportunista*, *opportuniste*, *opportunisti* (meaning "opportunist") in sports articles vs. all other articles. Taken together, these words seem to suggest that a negative *semantic prosody* (see, e.g., Sinclair (1998)) is associated with the word *opportunista/e/i* in non-sports contexts, as opposed to sports contexts. These findings confirm native speaker intuitions regarding the general language, according to which the word is indeed very negatively connotated, and hint at a specialized use in the sports jargon.

sports	*other than sports (selection)*	*other than sports (selection)*
buon	abile	infanticida
coraggioso	ambizioso	intrallazzatore
duttile	battagliero	maldestro
fromboliere	camaleonti	mediocri
furbo	certi	notori
grande	cinico	piccoli
grandioso	delinquente	politico
pronto	eterni	squallido
splendido	incallito	svenevoli

Table 4: Semantic prosodies: *opportunista/e/i*

Table 5 compares the use of loan words in 1987 and in 1992. These results were obtained by searching for words tagged LOA (see 2) in the respective corpora. Only the top 10 most frequent loan words from each year which are not attested in the other year are given here.

Table 6 compares sequences of words tagged as "modal"+ "main verb" in commentaries vs. news reports.

[4] The high accuracy achieved by this model is not particularly impressive, given that a "trivial rejector" (Sebastiani, 2002) facing the same task would achieve an accuracy of 94.52%.

[5] `http://www.natcorp.ox.ac.uk/sara/`

1987	*freq.*	*1992*	*freq.*
earning	24	ex-Urss	136
resettlement	19	fax	57
ears	11	disk	41
bulk	11	skinhead	34
section	10	lumbard	29
nuclear	9	outplacement	24
routier	9	annus	24
scoring	7	malus	20
ludens	7	rapper	20
review	7	core	20

Table 5: Diachronic perspective: Loans in ’87 vs. ’92

Only the top 10 most frequent sequences of words that fit this pattern are listed here.

comment	*freq.*	*news*	*freq.*
vuol dire	8449	potrebbe essere	696
può essere	8222	dovrebbe essere	585
potrebbe essere	6325	può essere	441
deve essere	5192	deve essere	314
può fare	4510	vuol dire	272
dovrebbe essere	4486	dovrebbero essere	205
può dire	4263	può fare	183
possono essere	2415	potrebbero essere	168
deve fare	2217	possono essere	163
doveva essere	1974	possa essere	161

Table 6: Structural patterns: MOD+VER bigrams in comments vs. news

6. Conclusion

As we said, the *La Repubblica* corpus is probably the largest freely accessible Italian corpus available to date. This does not mean of course that it is an “ideal” corpus of Italian, and indeed it has a number of weak points.

First and foremost, all its texts are instances of the same macro-genre – journalistic prose – and come from the same source. Second, the corpus has been processed and annotated automatically, therefore many typos, errors and idiosyncrasies remain. Third, extra-textual and structural information is scarce: It is not possible to search for two words appearing within the same paragraph, or for all texts appearing on the so-called “terza pagina” (third page, i.e., the page(s) traditionally devoted to culture, the arts etc.)

Yet this corpus also has some strong points, beyond its size. In particular, the current and prospected access methods are powerful and they allow for extreme flexibility. Also, the semi-automatic genre/topic categorization is a treat by itself, opening up numerous alleys for research (e.g., comparative studies of the language of sports vs. politics, of the rhetoric of comment vs. news articles. training domain-specific language models, etc.)

At the time of writing, work is underway to enlarge the corpus, add further annotation and improve accessibility.

To achieve the first goal, we are extracting the articles from the years 1993-2000. This is far from trivial, since there are differences in the way in which the data of each year were stored and formatted in the original CDs.

In terms of annotation, current efforts focus on lemmatization and morphological analysis. We expect that the morphological resources we are developing will also help improving the overall performance in POS tagging (along the lines of, e.g., Tamburini (2000)).

In terms of accessibility, we are working at making the fully annotated corpus available also via the SARA client, we are finalizing the web CQP interface, and we are experimenting with tools that allow interested users to retrieve large-scale quantitative data (a first step in this direction being the web interface to the cwb-scan-corpus command mentioned in 4. above).

7. References

L. Biagini, R. Bindi, S. Goggi, R. Marinelli, M. Monachini, P. Orsolini, E. Picchi, S. Rossi, N. Calzolari, and A. Zampolli. 2000. *Criteria and methods for building the Italian PAROLE corpus.* CNR-ILC Technical Report.

O. Christ. 1994. A modular and flexible architecture for an integrated corpus query system. *COMPLEX ’94.*

A. Finn and N. Kushmerick. 2003. Learning to classify documents according to genre. *IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis.*

G. Graffi. 1994 *Sintassi.* il Mulino.

T. Joachims. 1997. *Text categorization with support vector machines: Learning with many relevant features.* Technical report, Department of Computer Science, University of Dortmund.

T. Joachims. 1999. Making large-scale SVM learning practical. In B. Schoelkopf, C. Burges, and A. Smola, editors, *Advances in kernel methods – support vector learning*, MIT-Press.

M. Monachini. 1996. *ELM-IT: EAGLES specifications for Italian morphosyntax lexicon specification and classification guidelines.* EAGLES Technical Report.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. *EMNLP 2002*.

R. Rossini Favretti, F. Tamburini, and C. De Santis. 2002. CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. In A. Wilson, P. Rayson, and T. McEnery, editors, *A Rainbow of Corpora*, Lincom-Europa.

I. Schroeder. 2002. *A case study in part-of-speech tagging using the ICOPOST toolkit.* Technical report, Department of Computer Science, University of Hamburg.

F. Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. *International Conference on New Methods in Language Processing.*

J. Sinclair. 1998. The lexical item. In E. Weigand, editor, *Contrastive Lexical Semantics*, Benjamins.

F. Tamburini. 2000. Annotazione grammaticale e lemmatizzazione di corpora in italiano. In R. Rossini Favretti, editor, *Linguistica e informatica*, Bulzoni.

H. van Halteren, J. Zavrel, and W. Daelemans. 2001. Improving accuracy in word class tagging through the combination of machine learning systems. *Computational Linguistics*, 27:199–229.