Linguistic Miner: an Italian Linguistic Knowledge System

Eugenio Picchi , Maria Luigia Ceccotti, Sebastiana Cucurullo, Manuela Sassi, Eva Sassolini

ILC-CNR
Via Moruzzi, 1 - 56100 Pisa
[picchi, luigia, nella.cucurullo, manuela.sassi, eva.sassolini]@ilc.cnr.it

Abstract

Linguistic Miner is a project carried out at ILC whose objective is the development of an integrated system to build, organise and manage a corpus of Italian texts (of various origins and formats), and to design and constantly add new tools for the automatic extraction of tiered linguistic knowledge to be made available for many teaching, publishing, and other cultural purposes. The project is based on a notion that is preliminary to all the systems for corpus-based linguistic analysis: a language represented by the largest possible collection of heterogeneous texts is the best source of linguistic information at any level of analysis considered. The first goals of such a system are the semi-automated construction of an Italian data mine for the extraction of linguistic information, the validation of linguistic patterns, the installation of useful tools and resources for a range of different categories of Italian language users. The main feature of the project is its purpose of building large language reference corpora allowing for the creation and use of effective tools for the handling and processing, as well as the automatic linguistic synthesis, of such corpora.

1. Introduction

This study describes an ongoing project carried out by the *Istituto di Linguistica Computazionale* of Pisa. The project is called "Linguistic Miner" (LM). Its purpose is to create an integrated system of resources and tools to be used for the creation, management and use of a large collection of textual materials as a reference basis for studies, analyses and highlights on several linguistic issues concerning the Italian language. The project starts from the consideration that lies at the basis of all "corpus-based" linguistic analysis systems: the best source of linguistic information, at many different levels of analysis, is the language itself represented by the largest possible collection of texts of the most varied types. The largest the corpora available, the more heterogeneously they represent the different linguistic fields and the more they can be quantitatively and statistically analyzed and assessed, the more representative they will be of the linguistic world of a language.

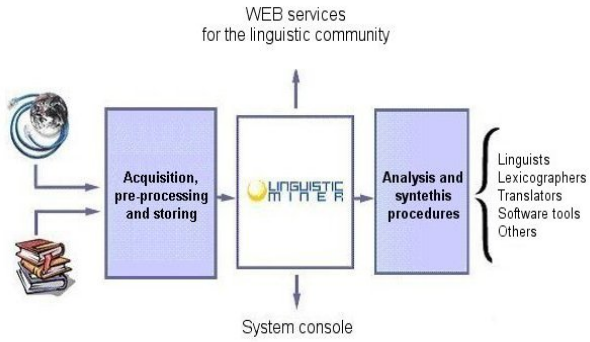

Figure 1: The structure of the LM project

The project draws its importance from the integration of its various constituents: the wealth and extreme variety of the textual mine add value and significance to the access and analysis tools, which, in their turn, allow for a better and more exhaustive exploitation of the mine (see Figure 1). The LM project originates from the availability of the textual processing and analysis tools provided by the "PiSystem" project and its base-engine DBT (Textual Data Base); it is from here that it takes its force to aggregate new ideas, methodologies, tools, objectives and – as we hope – new results that may be used by linguists in their task of assessing hypotheses, by lexicographers for a careful collection of useful information for lexicographic compilation, by language teachers in providing them with certification and verification tools, and finally by the developers of linguistics-based tools, such as automatic or assisted translation systems. The value does not rely on individual components, which are mainly already developed in the institute, but in their integration in a single flexible system which allows the user to select among a broad range of different functionalities.

2. Creation of the mine and further additions

The project consists of a really huge corpus of Italian texts to be continuously enlarged. We take up "quality" as a criterion for the selection of the text to be added to the LM, so that the texts will be considered as written published content or the like, even though the source is the Internet. We do not accept texts that are produced interactively such as, for example, mails, forums or blogs. The currently existing mine includes tested-quality materials, such as: databases, textual banks, newspaper archives that have been processed for a variety of different purposes and in several manners in the course of our activity. Such texts and similar contents will be added again and again, each time they become available. In the last few months, another source has been considered to continuously enlarge and enrich our corpus: the Internet, with the exponential growth of easy-access and low-cost web sites and contents available. Two different methods have been designed to prevent the danger of introducing doubtful quality texts into the mine: one intended for dynamic sites, whose contents are regularly added (from newspapers and magazines), and the other for substantially static sites, whose contents are not updated at predictable intervals.

For the dynamic sites, whose list is subject to change, an appropriate *spidering* procedure has been created to visit them at intervals corresponding to their dates of issue and to download only the new content to be automatically added in the text mine. These data retrieval operations are performed at night, when access to the web is easier and faster, with an automatic service module that extracts and converts the contents to be added into the mine.

Conversely, since prevalently static sites require checking, evaluating and subdividing texts into categories (at a first classification tier), a guided operational strategy has been adopted to allow operators to select web sites and subsets of web sites that are deemed to be significant and of sufficient quality. The download procedure is then started only for the items selected (see Figure 2).

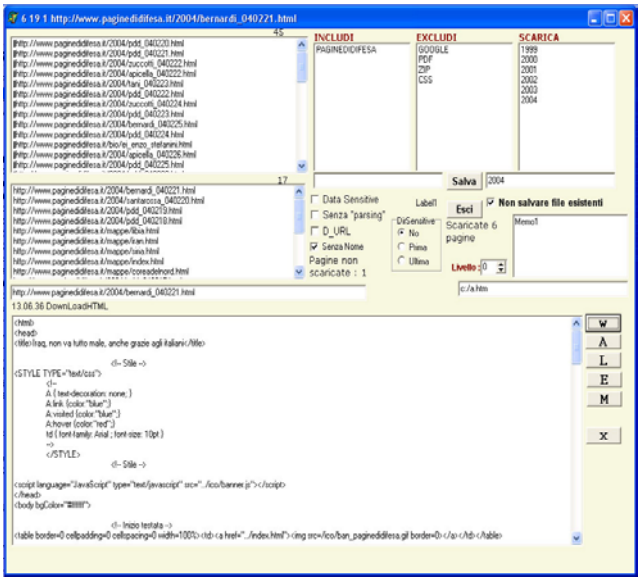

Figure 2: An example of spidering phase

2.1. The *copyright* issue

The first problem that is traditionally encountered in the construction of textual content corpora is copyright. The structure of the LM system, whose aim is to analyze and synthesize linguistic phenomena, does not include an information and document contents retrieval step, but rather a "de-structuring" of the textual materials in order to allow for search and extraction operations among the linguistic phenomena in the texts. The "de-structuring" phase is performed by an algorithm for the introduction of the contents into the textual data base, so that the contents may be accessed only by the parsing and exploiting procedures of the LM. Therefore, individual texts cannot be 'seen' in the mine, they cannot be read or reproduced, so that, in practice, it is impossible to know whether a text has been added into the mine or not. In this manner, the contents are no longer strictly dependent on their original source and, at the same time, using access functions of the LM system, the potential and objectives of the project remain intact in their being purely linguistic purposes, and not of content analysis.

The classification task of any text added to LM is performed before its insertion in the mine. After this operation, there is no longer a text but rather a field of contents – art, architecture, environment, medicine, etc. - where users can perform searches.

2.2. Parsing procedures and their relevant codification

Special parsing and transcodification procedures have been used to analyze original documents, to extract the textual component to be codified in the LM's text analysis procedure. These procedures have been mainly developed for HMTL format texts (see Figure 3) and similar tools have also been developed for other types of materials that have become available (Word, RTF, PDF, LIT). Specific functions, partly borrowed from the "PiSystem" project, have been applied to identify and process phenomena such as text structure, acronyms, abbreviations, proper names (single words or expressions), links to web sites and hypertexts, e-mail addresses, etc.

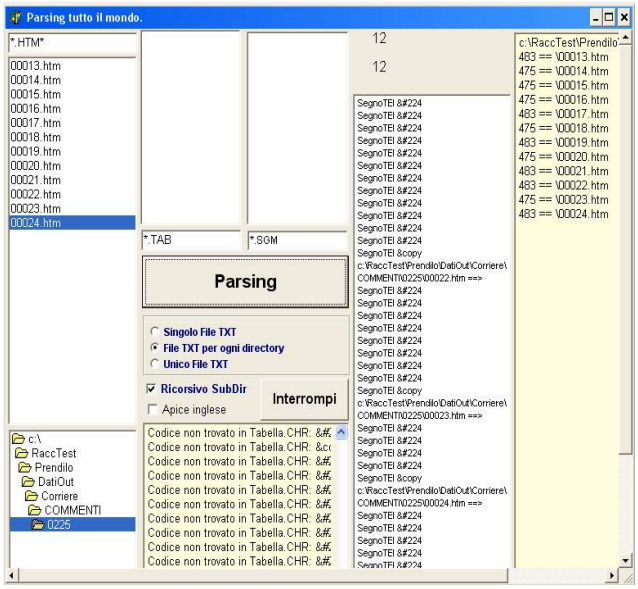

Figure 3: An example of HTML parsing session

2.3. The system console

The LM's system is controlled through a sort of console, through which the following operations can be performed:

- get mine status information,
- manage each section,
- add new contents,
- facilitate the classification work,
- select a subset,
- enable analysis procedures,
- manage all the intermediate analysis and synthesis steps performed in the mine.

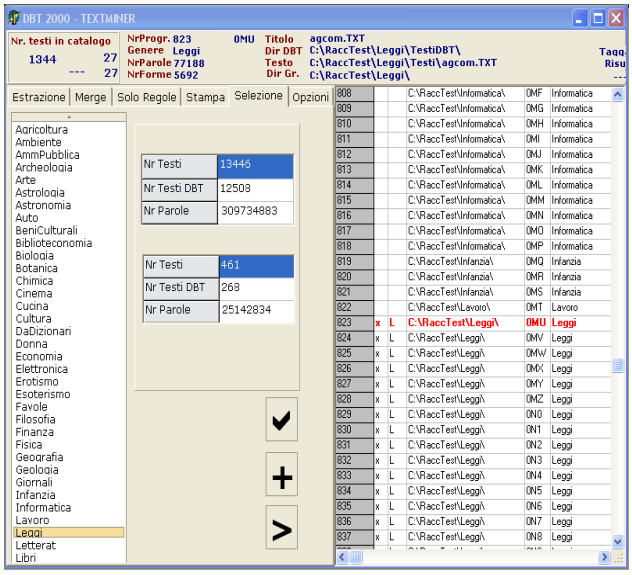

Figure 4: System console

Figure 4 shows a screen page of the system console where the selection function has been activated to view both total and partial quantitative data regarding the selected sector. A part of the material classified under category *'Leggi'* [Law] found in the mine can be seen.

3. Classification of individual texts

One of the most important features and purposes of the project is to classify each individual text that has been entered into the mine. This will allow the user to analyze, study and obtain the information and significant linguistic issues, subdivided by text typology, that describe language behaviour in different situations. The first tier classification consists in assigning a category to each text entered: the information required to perform this operation is drawn from the source of the material and by performing a first analysis producing a temporary classification. An automatic procedure is currently being developed to classify each individual text item of the mine; this procedure uses a statistical categorization approach that uses a reduced set of texts previously classified as "training corpus".

4. Automatic lemmatization

The next but not less important action is automatic lemmatization and POS-tagging, performed through the Pitagger procedure, a fundamental component of the PiSystem project. This procedure, which is developed based on a statistic approach, yields good results in the phase of disambiguation of morpho-syntactic analysis. The PiTagger reached a 97% of correctness and an error percentage is, however, unavoidable, even though, due to the law of large numbers - which is of fundamental importance in "corpus linguistics" – it will decrease as the corpus size is increased. We can verify this assumption in the results we can obtain from the LM: errors can be found but only with a few frequencies.

5. Multi-level textual data bank

At present, our research group is prevalently working at retrieving texts and transforming them into a huge textual data bank of the Italian language. The results obtained using the various analysis and linguistic information extraction functions will be an integral part of the project. Part of these functions have been already built and tested onto the available material, while others have to be 'invented' for new ideas or specific user requests. By doing so, the bulk of linguistic information will be constantly increased during the various analysis and synthesis steps, and made available for a variety of studies, researches and development of application systems, with the overall objective of creating an incredibly rich base of linguistic knowledge of the Italian language.

5.1. Analysis and use of the corpus

The phase of filing textual data, which is presently requiring the greatest efforts by the researchers involved, was followed by the analysis of data, a phase that will become increasingly important from now on. With the currently available tools, we can extract different frequencies and distributions of frequencies by form or lemma, left or right sorted word contexts, lemmas, locutions or groups of words, repeated sequences of words, statistic co-occurrences obtained through the application of *proximity indexes* (e.g., the *mutual information* index).

5.2. Identification of linguistic patterns

A recently developed linguistic investigation tool creates a finite state automaton, which allows linguistic patterns to be identified and applied to the whole mine or to a specifically selected subset. Specific models can be searched with this program, and linguistic formulas can be applied within the corpus (see examples in section 6).
In order for the corpus to be analyzed more effectively, the procedure executes the lemmatization phase automatically. This phase is executed just the first time the text is queried and the results are stored in the mine, thus remaining available for all subsequent searches.

6. Examples

To provide an example, we will now show the result obtained by applying a rule where a linguistic pattern is defined to retrieve substantives and their relevant adjectives. All pairs of substantives and adjectives are extracted for the selected sub-corpora: the rule has been applied to the texts of eight subsets: agriculture (Agr), environment (Amb), art (Art), biology (Bio), kitchen (Cuc), economy (Eco), XIX-XXth century literature (Let) and medicine (Med). After merging the partial results of each text with the merge procedure, the data have been processed for synthesis. For example, the term "ambiente" (environment) has been extrapolated from the general list of all substantives together with its adjectives.
Figure 5 shows the adjectives distribution of the hapax in each sub-corpus and clearly demonstrates how the use of adjectives offers a lexical richness in literature but always with low frequency values. Conversely, Table 1 shows all the pairs with a frequency greater than 1 in the different sub-corpora.

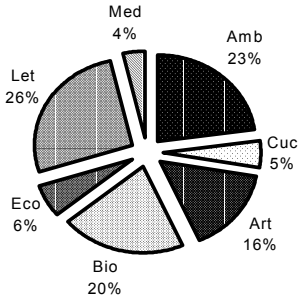

Figure 5: Graph of hapax adjectives

It is easy to guess the great potential of these tools applied to the whole mine of the LM project when the linguistic phenomenon analyzed is described in its entire complexity and richness verified in the reference corpus. A synthesis of the specific phenomenon available is therefore provided and made available for a wide variety of studies and applications. There are many categories of users who may be interested in such kinds of tools and resources: linguists who study the multifaceted aspects of languages, lexicographers who are by nature concerned with any sort of lexical and linguistic synthesis in general, or publishers, who may have a direct interest.
However, this first example of linguistic/lexical information extraction already provides enough evidence of the importance the LM can take up also for non-human users. In fact, in the case of an automatic translation system translating into Italian, each occurrence of a substantive-adjective pair offering multiple alternatives and combinations can be resolved by resorting to the bank of data extracted from the LM to select the most likely

solution based on the type of text to be translated.

Agr - 1.861.617		(Amb)		**Bio – 1.137.308**	
36	naturale	8	cosmopolita	19	specifico
14	urbano	8	acquatico	17	esterno
8	esterno	7	areato	7	mediterraneo
8	mediterraneo	7	croato	6	marino
8	rurale	7	significativo	6	particolare
7	confinato	7	indoor	6	interno
7	cosmopolita	6	economico	6	ricco
7	fresco	6	fresco	6	colturale
7	simulato	6	simulato	4	diverso
7	spazioso	6	artificiale	3	cosmopolita
6	marino	6	idoneo	3	protetto
6	costiero	6	abitativo	3	controllato
6	economico	6	alpino	3	fisico
6	fisico	6	industriale	3	nuovo
6	areato	6	detto	3	determinato
6	croato	6	eccezionale	2	abitativo
6	significativo	6	sanitaro	2	stesso
5	caldo	6	suggestivo	2	simile
4	idoneo	5	protetto		
3	chiuso	5	fisico	**Art - 1.734.792**	
3	scolastico	5	caldo	9	naturale
3	protetto	5	spazioso	8	domestico
3	controllato	4	acido	7	sociale
3	particolare	4	particolare	7	giapponese
3	aerato	4	diverso	6	urbano
3	sociale	4	britannico	6	familiare
3	alpino	4	cosiddetto	6	artistico
3	determinato	4	estremo	5	acquatico
3	diverso	3	scolastico	4	esterno
3	stesso	3	domestico	4	caldo
3	agricolo	3	controllato	4	artificiale
3	attuale	3	nuovo	4	simile
3	degradato	3	sociale	4	artificiale
3	estremo	3	determinato		
3	inospitale	3	ricco	**Eco - 1.163.928**	
3	isolato	3	stesso	14	esterno
3	mancino	3	umido	14	economico
3	rappresentativo	3	abitabile	7	naturale
3	regionale	3	adatto	7	ostile
3	ricevente	3	agricolo	6	concorrenziale
3	sardo	3	degradato	3	rurale
3	sterile	3	globale	3	simulato
3	sterilizzato	**Med - 1.730.254**		3	artificiale
3	ventilato	17	naturale	3	industriale
2	abitabile	14	scolastico		
2	cosiddetto	11	esterno	**Cuc - 1.662.965**	
2	globale	11	familiare	7	asciutto
2	acido	11	ospedaliero	6	fresco
2	britannico	8	acido	4	adatto
2	terrazzato	6	chiuso	3	particolare
Amb - 2.316.633		6	domestico		
111	naturale	6	protetto	**Let - 12.388.406**	
50	marino	6	controllato	14	piccolo
44	confinato	6	aerato	7	nuovo
28	idrico	6	umido	3	ostile
19	esterno	6	comune	3	vasto
19	costiero	6	malsano		
14	urbano	6	privo		
14	chiuso	6	pulito		
9	mediterraneo	5	specialistico		
9	rurale	4	specifico		

Table 1: Adjectives with *ambiente*

7. Perspectives

Our project is designed with the purpose of being constantly enriched and developed with the dynamic addition of textual materials in order to extend both the quality and the size of the mine's coverage including all accessible types of texts. It is our intention to continuously engineer the entire procedure in order to make it more user-friendly, as well as increasingly reliable. The main actions required to improve the global operation of the system should concern the following phases:

- pre-analysis, a step necessary to increase the modes and quality of classification of phenomena (such as any proper name classification), which should improve the subsequent analysis step;
- the classification method, in order to have more complex and significant grids for searching linguistic phenomena;
- synthesis and extraction of linguistic knowledge from the corpus, both by improving the pattern search procedure and by adding new tools for analysis, including upper level analysis tools.

The final, and rather ambitious, objective is to make the whole project available to users on the Internet with an "open source" approach and a perspective of enrichment and improvement of both textual material and parsing/exploiting tools. To this purpose, we are presently assessing the most appropriate means and functions, in the awareness that our initiative might also be enriched with the precious feedback from the real users of the system, whether they be the final users (linguists, translators, teachers, etc.) or the developers of analysis components to be applied and evaluated using the LM.

References

Berry, M.W. (Ed). (2004). *Survey of Text Mining: Clustering, Classification and Retrieval*, Springler-Verlag, New York.

Daille B. (1994). *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques*. Thèse de Doctorat en Informatique Fondamentale. Université Paris 7.

Daille B., Romary L. (Eds.) (2001). Linguistique de corpus, *Traitement Automatique des Langues* (TAL), 42(2), Hermès Sciences Publications.

McCallum A. K. (1996). *Bow: A toolkit for statistical language modeling*, text retrieval, classification and clustering. http://www.cs.cmu.edu/mccallum/bow.

Picchi E. (1994). Statistical Tools for Corpus Analysis: A Tagger and Lemmatizer for Italian, in *Euralex '94 Proceedings*, Martin W., Meijs W., Elsemiek M., Sterkenburg P., Vossen P. (Eds.), Free University of Amsterdam, The Netherlands.

Picchi E. (2003). PiSystem: sistemi integrati per l'analisi testuale, *Linguistica Computazionale*, Vol. XVIII-IX, special issue, Zampolli A., Calzolari N., Cignoni L. (Eds.), I.E.P.I., Pisa-Roma, pp. 597-627.

Picchi, E., (2003). Esperienze nel settore dell'analisi di corpora testuali: software e strumenti linguistici, *Informatica e Scienze Umane*, a cura di Marco Veneziani, Olschki Editore, Firenze, pp.129-155.

Picchi E., Ceccotti M.L., Cignoni L., Cucurullo S., Fiorentini G., Sassi M., Sassolini E., Turrini G., (2003). Linguistic Miner, in *Atti del Congresso Annuale AICA*, Trento, 15-17 Settembre 2003.

Sinclair J. (1991). *Corpus, Concordance, Collocation*, Oxford University Press, Oxford.

Zampolli A., Calzolari N., Picchi E. (1986). Italian Multifunctional Data Base, in *Standardization in Computerized Lexicography*, Proceedings, Leidoff V. (Ed.). Institut der Gesellschaft, Universität des Saarlandes, Saarbrücken, pp.69-86.