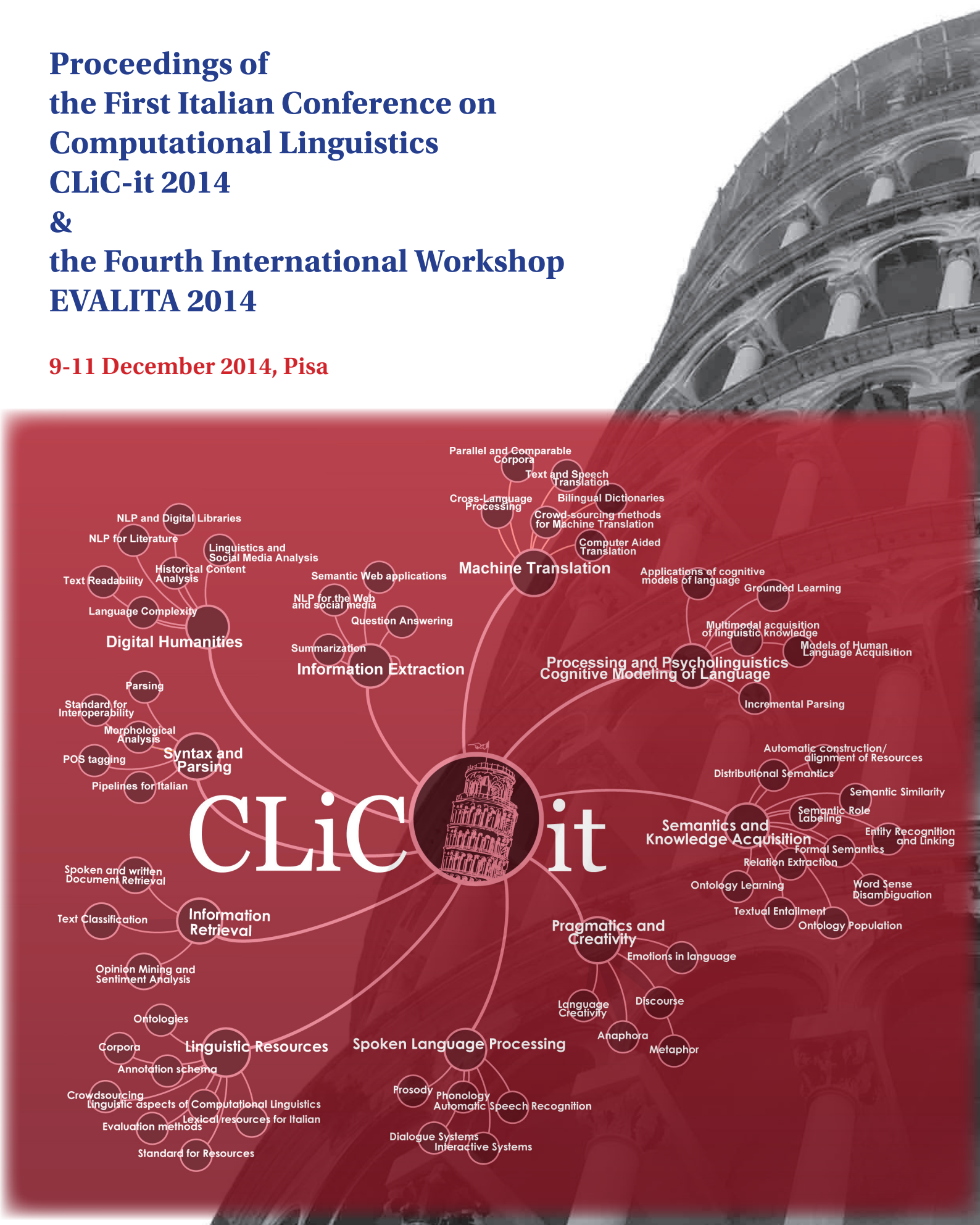
Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014

9-11 December 2014, Pisa



PISA UNIVERSITY PRESS

Volume I

The First Italian Conference on Computational Linguistics CLiC-it 2014

Proceedings

Editors
Roberto Basili, Alessandro Lenci, Bernardo Magnini

9-10 December 2014
Pisa, Italy

Società con socio unico Università di Pisa
Capitale Sociale Euro 20.000,00 i.v. - Partita IVA 02047370503
Sede legale: Lungarno Pacinotti 43/44 - 56126, Pisa
Tel. + 39 050 2212056 Fax + 39 050 2212945
e-mail: press@unipi.it
www.pisauniversitypress.it

ISBN 978-886741-472-7

10.12871/CLICIT2014164

Methods of textual archive preservation

Eva Sassolini
Istituto di Linguistica Computazionale "Antonio Zampolli"
eva.sassolini@ilc.cnr.it

Sebastiana Cucurullo
Istituto di Linguistica Computazionale "Antonio Zampolli"
nella.cucurullo@ilc.cnr.it

Manuela Sassi
Istituto di Linguistica Computazionale "Antonio Zampolli"
manuela.sassi@ilc.cnr.it

Abstract

English. Over its fifty-years of history the Institute for Computational Linguistics "Antonio Zampolli" (ILC) has stored a great many texts and corpora in various formats and record layouts. The consolidated experience in the acquisition, management and analysis of texts has allowed us to formulate a plan of recovery and long-term digital preservation of such texts. In this paper, we describe our approach and a specific case study in which we show the results of a real attempt of text recovery. The most important effort for us has been the study and comprehension of more or less complex specific data formats, almost always tied to an obsolete technology.

Italiano. *L'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) nella sua storia cinquantennale ha accumulato una grande quantità di testi e corpora che sono stati conservati in vari formati e tracciati record. L'esperienza storica nell'acquisizione, gestione e analisi del testo ci ha permesso di formulare un piano di recupero e conservazione digitale a lungo termine di materiali testuali. In questo articolo, descriviamo il nostro approccio e un caso di studio specifico in cui sono riportati i risultati di una reale operazione di recupero. Il maggiore impegno è stato dedicato alla comprensione di particolari specifiche di formato più o meno complesse, ma quasi sempre legate ad obsolescenza tecnologica.*

1 Introduction

The international scientific communities consider electronic resources as a central part of cultural and intellectual heritage. Many institutions are involved in international initiatives[1] directed to the preservation of digital materials. The Digital Preservation Europe project (DPE) is an example of "best practice" realization. The principal issues concern the techniques and processes of digital memory management, but also of concerted action at both the national and international levels. *Digital preservation* is certainly a very challenging task for individual institutions.

The means of *digital preservation* can be explained by the following definition: "Digital preservation combines policies, strategies and actions to ensure access to reformatted and born digital content regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time." (ALA 2007:2)

In our specific case we are engaged in looking for systems and techniques necessary for the long-term management of digital textual materials, stored at ILC over many years of work. At the beginning we did not know the age of all the materials. However, at a later stage, we found a number of digital materials, some of which dated as far back as the 70's. For this reason, the work of recovery is extremely complex and demanding. Firstly, the format of file/text encoding was often obsolete and not associated with exhaustive documentation. Secondly, many texts contained disused linguistic annotation schemas. Conse-

[1] Digital Preservation Europe (DPE) was a collaborative European digital preservation project that ran from 2006 to 2009, aimed at creating collaborations and synergies among many existing national initiatives across the European Research Area.

quently, we have adopted different strategies for "textual resource" retrieval, before focusing our attention on the application of standardized measures for conservation of the texts.

2 Text analysis

A *format specification*, when available, provides the details necessary to build a file from a text, and it establishes the admitted encodings and software applications able to decode the file and make its contents accessible. These documents can be of extremely variable size depending on the complexity of the format. However, the file *format specification* has not always evolved with the related software. Obsolete software and file formats, as well as storage medium, are today open issues.

A file format may become obsolete for several reasons:

- the latest software versions do not support the previous files;
- the format itself is superseded by a new one, or becomes more complex;
- the format is not so widely adopted, or the scientific community does not support the creation of compatible software;
- the format is no longer compatible with the current computers;
- the software supporting the format has declined.

Digital formats are a challenge for text conservation. In the early decades of computing, only few people were aware of the threat posed by the obsolescence of file formats for long-term digital preservation. A systematic effort for collecting software documentation or all the specifications necessary for the conservation of textual files was missing. With no proper documentation, the task of interpreting the contents of an old file is very demanding. It is only recently that we have started to catalogue, document, and understand these contents, together with their relationships and variations.

While most of the software is regularly updated, the relevant files become sometimes obsolete and therefore unable to meet the new format requirements, thus making even the latest versions of the software unreadable. Moreover, if the older versions are no longer available, or do not run on a recent computer or in the current version of the operating system, the data is lost. Owing to the complexity and nature of some file formats, it can be extremely complex to know whether a converted file in another format has retained all its features.

2.1 Conservation measures

Preserving the information should be the main goal. It is the information content of a document (tokens, linguistic annotations, critical apparatus, figures, etc.) that should be maintained in compliance with international standards. The standards usually need to respond to a large community of users, not linked to individual economic interests[2].

However, compatibility with the standards available is not generally priority for data producers, because either it is costly, or because there are commercial pressures that render the older formats quickly obsolete.

On the other hand, standard formats are not necessarily the best choice for all situations, but they offer great advantages for long-term preservation and storage. Finally, to reduce the risk of obsolescence a standardization process is required, which should primarily concern the formats at greatest risk, like the ones created by obsolete or outdated software versions.

3 First texts in electronic format

Electronic processing has always been articulated into three basic steps: input (input or acquisition of data within the procedure), processing (analysis of the material and specific processing depending on the intended purpose) and output (output, on suitable media, of the results of the previous stages). The output of any processing phase may be considered as a final result in its own right, even if - in a specific project - it can be an intermediate analysis subject to subsequent processing phases.

A fundamental parameter for the whole process is the type of storage medium used to preserve the material at the different processing stages. In the past, the only one choice available was the magnetic tape, which required sequential access to the data: in order to read (or write) a piece of data recorded on that medium it was necessary to read (or rewrite) all the preceding data in sequential order. This technology entailed objective

[2] The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its principal deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. (`www.tei-c.org`)

limitations concerning the possibility of implementing particularly sophisticated and efficient algorithms for data access.

3.1 Description of textual materials

During the analyses of textual materials we identified various levels of recovery in compliance with international recommendations, and at the same time to archive a standardized and universally recognized format that would allow the exchange and storage of materials. For the data to be better understood, we must explain the procedures which have produced them. For many years the "input" phases have only been possible through the preparation of punched cards that seem to belong to the pre-history of computer science. Units of data entry have come later, able to record on magnetic media or to operate in direct connection with the electronic computer.
The "output" phases normally consisted in the creation of two types of results:

- storing the results both as recovery of the same results, and as intermediate input to further processing;
- printing the results on fanfold paper to obtain the final output of the process and drafts used to check the correctness of the work performed, or to dispose of a working medium that can be enriched with new data or used to classify the previously stored data.

4 Main text problems

The first problem of management of a text concerns character encoding, which included sets of diacritics, or languages with non-Latin alphabets, or texts with multiple levels of annotation (e.g. comments, notes in margin, various types of footnotes, structured or dramatic text, etc.).
In the past the "ANSI format" (characters encoding belonging to the ISO 8859 family) generally represented the standard, with all the problems related to the sharing of sets of positions between the tables of the ISO 8859 family. Today the development of new encoding standards at the international level imposes shared models of representation of the data: XML TEI and UNICODE encoding.

4.1 Text acquisition strategy

If we retrace the steps of text acquisition for which ILC was among the pioneers in the industry, we see that there is no single conversion mapping, but that it is necessary to assess different types of material and their specific recovery paths.

At present, it is possible to make only an estimate about textual heritage. However, this is sufficient to set up a common procedure and useful to evaluate the costs of the entire operation. Depending on the types of material (from texts on magnetic tapes to machine readable and editable digital texts) we have hypothesized different phases of recovery. Therefore it is impossible to define a series of procedures valid for all types of contents or data.

We cannot forget the software DBT[3] that has often been used for the treatment of ILC texts.

For example at least three phases are required in order to convert a text file with obsolete character encodings: a first mapping involves the conversion into an intermediate format, typically an ANSI encoding; a second format is produced by the recognition, management and remapping of all the annotations inserted in the text; finally, the last phase involves the construction of a parser that can read these annotations and convert them into appropriate TEI-XML tags.

Source text	Perc.	Transition phases (TP) required	Meta data
Text on magnetic tape	10%	Many TP type	study and research in the ILC archives
Text divided into separate resources	5%	TP>3	recovered from paper-based data
Text in obsolete file	10%	TP>2	recovered from paper-based data
Digital text with obsolete character encoding	10%	2<TP<3	recovered from paper-based data / digital format
Digital text	65%	One TP	recovered from the digital format

Table 1: acquisition strategy

A more complex case is represented by lemmatized texts, where the annotations are at the level of words and then become more extensive. Even for the annotation of lemmatized texts there has been a wide use of the DBT software. In the acquisition protocol for this type of text, this level of analysis is added to the others together with the evaluation of the type of software tool that was used at the time.

[3] DBT (Data Base Testuale, Textual Data Base) is a specific software for the management and analysis of textual and lexical material.

Source text	Transition phases (TP) required	specific annotations type encoding	Meta data
Texts on magnetic tape	Many type TP	?	long and difficult work
Text divided into separate resources	TP>3	DBT type encoding	recovered from paper-based data
Text in obsolete file	TP>2	Obsolete type encoding	recovered from paper-based data
Digital text but obsolete character encoding	2<TP<3	Specific type encoding	recovered from paper-based data / digital format
Digital text	One TP	ILC text encoding	recovered from digital format

Table 2: annotated text acquisition strategy

5 Results

A concrete example of application of the procedure for recovery of texts belonging to the heritage of texts of the Institute (briefly "ILC Archives") is related to the work resulting from a scientific agreement between ILC and the "*Accademia della Crusca"* of Florence. The researchers of the *Accademia* were especially interested in recovering the lemmatized corpus of "*Periodici Milanesi"*[4].

The archive dates back to the early 80's and originates as post-elaboration of the lemmatization procedure implemented by ILC researchers and technicians in the 70's. The output format consists in files made up of fixed fields, each containing several types of information. The first challenge consisted in interpreting and decoding both the file format and the complex annotation scheme.

The most complex part of the decoding of the "starting-point" files (in TCL format/ASCII) concerned the retrieval of text and related annotations: lemma, POS (part of speech) and any other semantic type of information. For a correct interpretation of the data records contained in the lemmatized texts, a preliminary study phase was made. An example is shown in the figure below.

```
T1 162468  0404□4   per
L1per                          per
T3 162469  1212□4   inserirvelo
L1inserire                     inserire
L2egli                         egli
L3vi                           vi
T01162470  0101 180'
```

[4] Digital materials extracted from "La stampa periodica milanese della prima metà dell'Ottocento: testi e concordanze", edited by Giardini (Pisa, 1983), authors: Stefania De Stefanis Ciccone, Ilaria Bonomi, Andrea Masini. Management of the text required the advice of Eugenio Picchi and Remo Bindi.

The fragment of original text encoding shows the complex representation of the format used. As a matter of fact, the information is expressed by a complex annotation scheme, whose interpretation and decoding represented the first phase of work.

The complexity of the original format required a conversion in two steps:

- a first step in which the texts were converted from the original format to a DBT-like format to favor a simple check on the correct decoding of the source format with no loss of information;
- a second step required the representation of the text in XMT TEI with Unicode encoding.

The archive of "*Periodici Milanesi*" contains a collection of 58 newspapers (1800-1847), organized in seven main categories: Political Information, Literary Magazines, Magazines varieties, Technical journals, Magazines theater, Almanacs, Strennas.

Corpus analysis and results:

- 879,129 tokens;
- 59,639 different forms;
- a TEI P5 XML file for each article of the corpus (2277 files), where all lemmas are appropriately coded.

Extraction of the main linguistic features:

- index of 975 spelling variants of the words;
- index of 312 different multi-words in the corpus;
- list of 710 Latin and French forms that have been coded as "foreign words".

6 Conclusion

The preservation of that data produced with outdated technologies should be handled especially by public institutions, as this is part of the historical heritage. Therefore, it is necessary for us to complete this work, so that the resources can be reused. This will be possible only through a joint effort of the institutions involved at the regional, national and international levels. ILC is currently establishing a number of co-operation agreements like the one with the "*Accademia della Crusca*", in an attempt to gather data resources for maintenance, preservation and re-use by third parties.

References

Alessandra Cinini, Sebastiana Cucurullo, Paolo Picchi, Manuela Sassi, Eva Sassolini, Stefano Sbrulli. 2013. *I testi antichi: un patrimonio culturale da conservare e riutilizzare*. In: 27a DIDAMATICA 2013, Tecnologie e Metodi per la Didattica del Futuro, Pisa. (Pisa, 7-8-9 may 2013). Proceedings, AICA.867-870.

Eugenio Picchi, Maria L. Ceccotti, Sebastiana Cucurullo, Manuela Sassi, Eva Sassolini. 2004. *Linguistic Miner. An Italian Linguistic Knowledge System* In: LREC Fourth International Conference on Language Resources and Evaluation (Lisboa-Portugal, 26-27-28 may 2004). Proceedings, M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa, R. Silvia (eds.).1811–1814.

Eugenio Picchi. 2003. *Pisystem: sistemi integrati per l'analisi testuale*. In Linguistica Computazionale, Vol. XVIII-IX, I.L.C. and Computational Linguistics, special issue, A. Zampolli, N. Calzolari, L. Cignoni, (Eds.), I.E.P.I., Pisa-Roma. 2003.597-627

Eugenio Picchi. 2003. *Esperienze nel settore dell'analisi di corpora testuali: software e strumenti linguistici, Informatica e Scienze Umane*. In LEO (Lessico Intellettuale Europeo), a cura di Marco Veneziani, S. Olschki editore Maggio 2003. 129-155

ALA (American Library Association). 2007. *Definitions of digital preservation.* Chicago: American Library Association. Available at: `http://www.ala.org/ala/mgrps/divs/alcts/resources/preserv/defdigpres0408.pdf`

Ingeborg Verheul. 2006. *Networking for Digital Preservation: Current Practice in 15 National Libraries.* Munchen: K.G. Saur 2006.