A Tool-box for Lexicographers

Claudio Giuliano
ITC-Irst
Via Sommarive 18, I-38050 Povo - Trento, Italy
giuliano@itc.it

Abstract

The Tool-box for lexicographers is a web-based application for accessing and updating lexical resources. The core of the system is a scalable object-oriented Java program that extends web server capabilities using Servlet technology. The client side - a standard web browser - allows the user to query multiple lexical resources in a uniform way. Since the legacy data format was strongly irregular both in structure and encoding, XML has been adopted as a common logic data format. A number of scripts have been implemented to make the encoding format uniform and convert the data into XML. The lexical resources currently made available through the web interface are dictionaries and corpora. The former can be queried by a Boolean combination of keywords, while the XML structure can be used to constrain matches to certain labels. The latter can be used to retrieve concordance lines and to investigate collocations. At present, the system is under integration and testing by a team of language professionals involved in the developing.

1 Introduction

The Tool-Box for lexicographers has been developed within the EU-funded project TALES[1] (Automatic Treatment of the Languages Ladino and Sardo). Started in November 1999 and still running, the project addresses the need for a suitable computational infrastructure for studying Ladino and Sardo, two minority languages spoken in Italy.

The need of working simultaneously on large corpora and multiple dictionaries made the software applications used by the language professionals involved in this project inadequate for many reasons. First, to carry out the same task on different lexical resources they had to use different applications. For example, searching for the same word in different dictionaries required repeating the same search in each dictionary, and if the word also had to be searched through corpora, another application had to be used. Second, the lexical resources were available in a wide variety of formats, not always suitable for the application employed. Finally, whereas corpora are relatively stable, dictionaries change continuously, and thus distributing updated copies of the dictionaries to the team members became a problem, especially when more than one person was working on the same resource. From the scenario described above two specific goals emerged: the conversion of the available lexical resources into a non-proprietary standard format, and the need for an application for accessing and updating lexical resources through a uniform interface.

In the last few years a large number of software tools have been developed to access and update corpora and dictionaries (see [Luz 2000; Rychlý 2000; Armstrong et al. 2000]), but the difficulty in adapting these tools to fit the project needs required a reimplementation for a new application. Therefore, the Tool-box is a tailor-made system - on the basis of the available resources - that groups together much of the functionality of other systems. Nevertheless, the resulting application is not a highly specialised program as one might have

expected, since the software has been developed with the goal of reusability in mind [Veronis and Ide 1996].

2 Data conversion and indexing

The available lexical data are of two types: a set of bilingual (Italian-Ladino, Italian-Sardo) and trilingual (Italian-German-Ladino) dictionaries stored in HyperCard and FileMaker format; two monolingual corpora (Ladino and Sardo) and a bilingual corpus (Ladino-Italian), all in different file formats, such as HTML, Microsoft Word and plain text.

Since the legacy data format was strongly irregular both in structure and encoding, XML has been adopted as common logic data format. The process of exporting the dictionaries to XML required two phases: first, classes of dictionaries with a common structure were detected and for each of them a specific DTD was designed; second, specific scripts were implemented to convert the dictionaries into XML according to the DTDs. On the other hand, the DTD defined in XCES[2] [Ide et al. 2000] was adopted to encode the bilingual corpus, while the monolingual corpora were available in plain text[3].

An inverted index - based on B-tree files - was implemented to improve keyword search through dictionaries and corpora. Indexing is performed off-line and can be pursued incrementally. At the moment the file formats handled properly by the indexer are plain text, HTML and XML. A list of stopwords and a minimum word length can be provided to reduce the index size.

In order to improve system performance the dictionaries are stored in a relational database. To map XML schema to relational model an object-relational mapping has been devised[4]. Since the structure of the available dictionaries is inherently data-centric, the mapping works well. Data records can be imported and exported into XML format. A middle-layer converts database records to and from XML hiding the underling database to the other modules.

The tools developed for the data conversion and indexing are part of the system, but at the present they are not available via the web interface; therefore, the database can only be updated off-line.

3 The Tool-box architecture

The Tool-box is a client-server application that allows a user to access lexical resources in a uniform way through a web-based interface. The following subsections briefly describe both the server side and client side.

3.1 Server side

The server side is an object-oriented program written in Java that extends web server capabilities using Servlet technology. Clients connect to the server using the HTTP protocol, a front-end object interprets the requests and delegates to the back-end objects the task of retrieving the data and building the query result; corpora and dictionaries manipulations are handled by proper back-end objects that can be easily extended to add new functionality. The query result is built in XML format and converted into HTML using an XSLT processor

[Clark 1999], which applies external CSS or XSL style sheets to obtain different presentations of the same data, depending on the user's profile.

All time-consuming tasks are performed on the server, while the client is responsible only for the presentation of the query result. This enables users with a slow internet connection to access the lexical resources properly. Query results are divided into multiple pages and cached on the server.

3.2 Client side

Any web browser supporting dynamic HTML can be used as a Tool-box client. This approach guarantees both portability and intuitive access to the system, as most users are accustomed to web pages. The web interface has been designed to allow users to access several lexical resources available in a uniform way, hiding data format and implementation details on the server side.

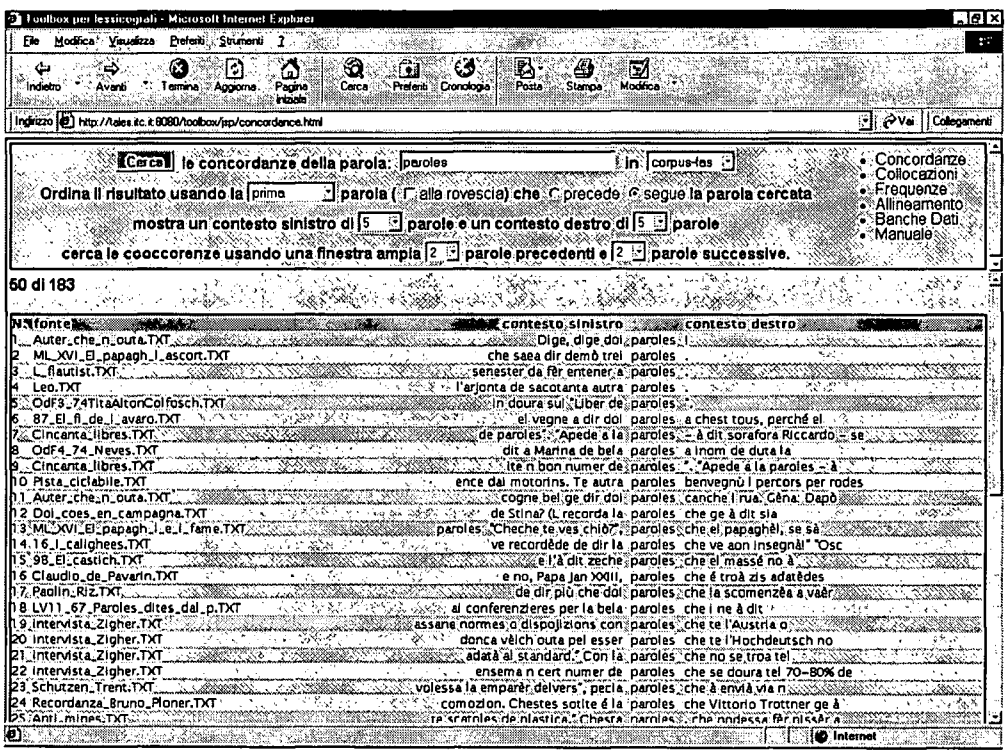

Figure 1: Accessing corpora. The top frame shows the query pane. The bottom frame shows the query result; the concordance lines are sorted according to the first word following the keyword.

The key aspect of this system is a mechanism for parallel querying of multiple dictionaries and corpora. Data are searchable by a Boolean combination of keywords, including regular expressions. Moreover, the XML structure can be used to constrain matches to certain labels.

For example, searching for a word can be limited to the examples and phraseology of the dictionary.

The Tool-box allows the retrieval of concordance lines from a corpus and the investigation of collocations. The concordance output is shown in KWIC format. The query has a high level of configurability, as several parameters can be arranged to make analysing the concordance lines easier. For example, keyword context can be enlarged and the concordance lines can be sorted by any word belonging to the keyword context. Statistical calculations against the set of concordance lines found can be performed in order to discover collocations, and different scoring evaluation functions are available. In particular, for each collocation candidate the following statistics are calculated: frequency, mean, variance, *t* test, Pearson's chi-square test and likelihood ratios. Finally, word frequency lists are calculated off-line and are accessible as static HTML pages, and concordance lines can be obtained directly by clicking on the word forms.

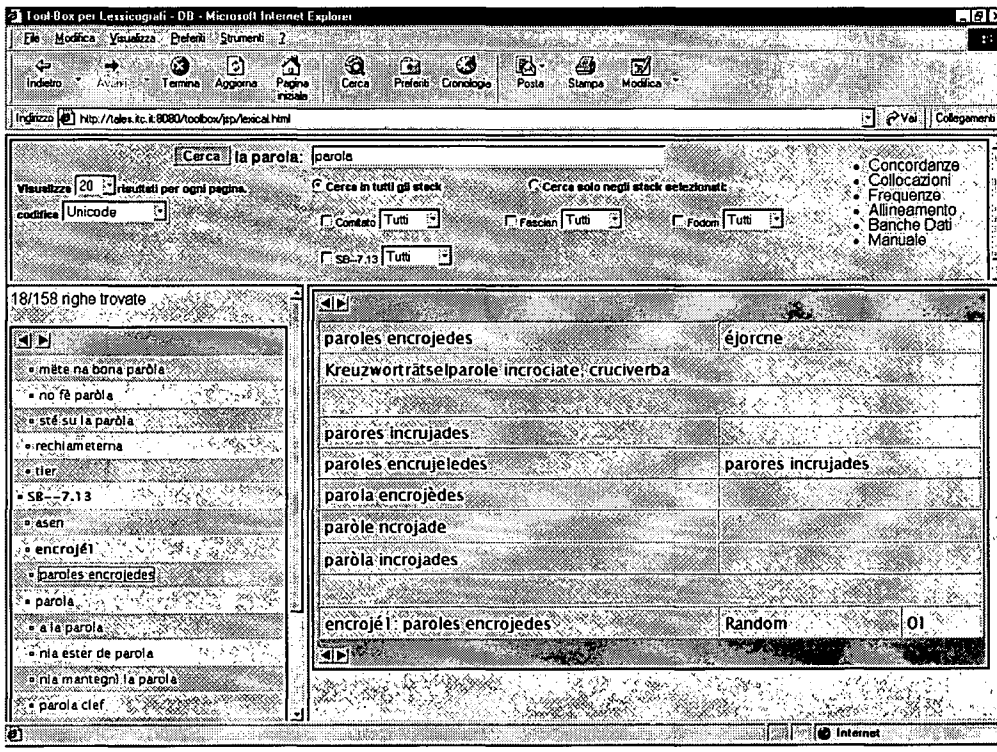

Figure 2: Accessing multiple dictionaries. The top frame shows the query pane; parallel queries on the available dictionaries can be run. The bottom left pane shows the entries that match the query; the entries are grouped hierarchically to facilitate searches. The bottom right pane shows the selected entry.

4 Conclusion

A web-based application for accessing and updating multiple lexical resources has been provided to the project members. The client-server implementation allows an up-to-date view of the available resources to be maintained easily and to apply an access control through a login policy. The web interface allows both portability and intuitive access to the services provided, and fast integration with other web-based services. The use of Java and HTML makes the application portable on different platforms, while the use of XML guarantees a clear separation between content and layout, giving the system flexibility and maintainability. The Tool-box is intended not only for lexicographers or terminologists, but also for professional translators, journalists, people working in the public administration, and everyone who has to write in two languages - Ladino and Sardo - that have a limited availability of lexical resources. A short term goal is to allow database updating through the web interface, and to improve the overall system performance.

5 Endnotes

[1] The TALES web site is located at http://tales.itc.it/.

[2] Refer to http://xml-ces.org for more information on XCES.

[3] A basic annotation process is planned for a further phase of the project.

[4] Refer to http://www.rpbourret.com/xmldbms/index.htm for more information about the object-relational mapping.

6 References

[Armstrong et al. 2000] Armstrong, S., Brace, C., Petitpierre, D., Robert, G., Walker, D. 2000. DicoPro: An Online Dictionary Consultation Tool for Language Professionals, in Ulrich Heid et al. (eds.) *Proceedings of the Ninth Euralex International Congress, Euralex 2000, vol I*, Universität Stuttgart, Germany.

[Clark 1999] Clark, J. (ed.), 1999. XSL Transformations (XSLT). Version 1.0. W3C Recommendation. http://www.w3.org/TR/xslt.

[Ide et al. 2000] Ide, N., Bonhomme, P., Romary, L., 2000. XCES: An XML-based Encoding Standard for Linguistic Corpora, in: *Proceedings of the Second International Language Resources and Evaluation Conference*. Paris: European Language Resources Association.

[Luz 2000] Luz, S. 2000. A Software Toolkit for Sharing and Accessing Corpora Over the Internet, in: *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000*, 1749-1754, Athens, Greece.

[Rychlý 2000] Rychlý, P. 2000. GCQP - Multiplatform Graphical User Interface to the CQP corpus manager, in: Ulrich Heid et al. (eds.) *Proceedings of the Ninth Euralex International Congress, Euralex 2000, vol I*, Universität Stuttgart, Germany.

[Veronis and Ide 1996] Veronis, J., Ide, N., 1996. Consideration for the Reusability of Linguistic Software available at http://www.lpl.univ-aix.fr/projects/multext/LSD/LSD1.html