THE OLISSIPO AND LECTIO PROJECTS

Giuseppe Cappelli, ILC – CNR Pisa Italy e-mail: beppe@ilc.cnr.it
Paulo Alberto, CEC – University of Lisbon e-mail: paulo.alberto@fl.ul.pt

OLISSIPO (Omnis Latinitatis Instrumentum Secundum Scholarum Instructionis Propositum Ordinatum) is a prototype developed by the Centro de Estudos Clássicos of the University of Lisbon and the Istituto di Linguistica Computazionale of CNR in Pisa, thanks to a common research project in the framework of the scientific agreement between the Consiglio Nazionale delle Ricerche (CNR – Italy) and the Gabinete de Relações Internacionais da Ciência e Ensino Superior (GRICES - Portugal).
OLISSIPO extracts lists of basic vocabulary from any Latin text and displays them together with linguistic and extra-linguistic information stored in a database. It contains other functionalities, such as statistical analysis, search of words in the text, displaying of the context, search of words in the database (which can be used as a dictionary).
The LECTIO project has also been carried out by the Centro de Estudos Clássicos of the University of Lisbon and the Istituto di Linguistica Computazionale of CNR in Pisa. Designed to be a follow-up of OLISSIPO, it aims at producing a prototype for analysis of Latin texts with more functionalities. The final result will be an open tool to be used both in teaching/learning and in scientific research.

1. Introduction

One of the problems to be dealt with in teaching and learning Latin is the student's low level of knowledge of vocabulary. It is fundamental that students progressively acquire and consolidate a domain of Latin vocabulary that is sufficient and appropriate for understanding the text they are studying, starting with the basic structure rules and interdependence of the words in the sentences.

In fact, the Latin teacher's experience shows that very often the students present considerable gaps in vocabulary which prevent them from acquiring a reasonable understanding of the Latin text. As well as having insufficient lexical knowledge, students are frequently incapable of establishing the relationship between the vocabulary, even between simple and compound, and comparing the Latin lexicon with that of their mother tongue.

The learning experience of some students has highlighted the need for a data processing tool which, when applied to a Latin text chosen by the teacher or the student, automatically produces results aiming at the presence of basic vocabulary**,** morphological categories and basic elements of the sentence. This would make it easier for the teacher to choose appropriate texts for his course. Furthermore, a tool capable of providing additional data concerning the lexicon (relationship with other words, reference to the phenomenon of composition and origin, dependence on Portuguese vocabulary) would also be extremely useful for the students.

The opportunity to develop an application focusing on these issues came about in 1996**,** upon the invitation of Professor Antonio Zampolli, director of the Institute of Computational Linguistics in Pisa (CNR). He was an researcher with vast experience and humanistic talent who, even at the beginning of language processing and with the help of a computer, worked to Seneca's concordances (1975, with R Busa) and, years later, to Sìmaco's concordances (1983, with V Lomato and N Marinone). Contacts were established between his prestigious Italian institute and the Centro de Estudos Clássicos da Universidade de Lisboa The aim was to work together on a common project within the field of the Portugal-Italy Scientific Working Convention This is how the OLISSIPO project came about.

2. Description of OLISSIPO

OLISSIPO can be described as a working environment in which the user is provided with different functions enabling him to work on a selected text, consult the analysis results and useful statistics, look for a specific work in the text, and modify the classification assigned to each word.

Furthermore, one specific function allows the user to access the data base, designed for classifying the words. Having processed the text**,** the user has at his disposal a function for extracting/choosing the lemma from the processed text and constructing/updating the basic vocabulary. OLISSIPO's objective is to create basic vocabulary lists and this can be achieved with individual strategies. Depending on the students' preparation and the selection of texts to be submitted to OLISSIPO, the teacher defines the most appropriate strategy (content and timing) for constructing the basic vocabulary.

The interface and other functions of OLISSIPO has been written in Visual Basic. For the tagging/lemmatisation programme (MORPH) and the disambiguation (DISAMB) programme the programming language C++ has been used This means that more in depth processing can be carried out in and in less time. MORPH reads the file containing the text to be processed, determines the occurrences and then consults the dictionary indicated by the user. If the word under examination is found in the dictionary**,** the classification (or classifications in the case of homography) is taken from the dictionary and associated to the word.

If the search provides a negative outcome the word is segmented to the right for eventual enclitic identification. If an enclitic is present, by using the word without the enclitic as a search key, the dictionary is consulted again**,** and if the search is positive the word is classified with the morphological information, and associated to the individualised enclitic.

One specific module of MORPH automatically identifies and classifies proper nouns. A word is recognised as a proper noun if it starts with a capital letter and is not preceded by strong punctuation. The morphological label is taken from an open list in which classification corresponds to each ending.

Words that are not analysed by MORPH are processed by LEMLAT, a computerised lemmatisation programme for Latin language developed at the Institute of Computational Linguistics in Pisa.

The DISAMB programme reads the file containing the labelled text, associates to each word the Part of Speech and the morphological information (case, gender and number) and extracts a sentence defined by strong punctuation. It then skims over the sentences and when it finds a homographic word (with more than one lemma or only one with several morphological values) it applies several morphosyntactic disambiguation rules for selecting the correct classification.

The interface written in Visual Basic has been studied in order to help the user with little experience in using computer tools for linguistic analysis.

When the user runs the OLISSIPO programme the window shown in figure 1 appears. In the bottom right three flags indicate the language the user wishes to interact with in the programme. To select the desired language the user simply clicks on the flag.

Figura 1

Apart from Italian and Portuguese, the languages of the Project partners, English is also available. Having clicked on the appropriate flag the window shown in figure 2 is opened which contains buttons for activating the OLISSIPO functions.

Figura 2

The icons on the buttons represent the eight OLISSIPO functions listed below.

1. *seleziona testo*: to select the text to be analysed;
2. *analisi*: for carrying out processing of the selected text;
3. *risultati*: to display the processing results;
4. *ricerca per forma*: to carry out searches within the text using the form as the keyword;
5. *ricerca per lemma*: to carry out searches within the text using the lemma as the keyword;
6. *statistiche*: displays the simple statistics provided by the programme;
7. *lessico*: to modify the information contained in the database;
8. *vocabolario basico*: to update the basic vocabulary after processing of a new text.

Only one window is shown below which relates to the analyses. For others please refer to the article *OLISSIPO – entre filologia e informatica: recurso para gerir o estudio do texto latino*, (**Euphrosyne**, 2004).

figura 3

The processing of the text is guided by several parameters that the user can configure according to his needs:

1) The LEMLAT[1] morphological analysis programme can be used interactively. This is helpful for checking the classification of a specific word or for carrying out didactic demonstrations;
2) The results can be displayed on the screen for initial checking;
3) It is possible to ask the programme to save the result of the processing onto a file. This is obligatory if the user wants to use the analysis functions offered by the programme;
4) It is possible to tell the programme to substitute the reference dictionary, used for automatic labelling, with one of its own which is more suitable for the analysis to be carried out.

Occasionally, when there has been a prototype presentation, as in the *Colloquium Didacticum Classicum Olisiponense XVII* (Lisbona, 30 Settembre – 3 Ottobre 1998), at the meeting of the Bureau International of the Didactique des Langues Anciennes (Gand, 11-12 Maggio 2000) and in the *Jornada Cientifica – Estudos Clássicos e Nova Filologia* (Lisboa, 17-18 Maggio 2001), new functions have been suggested. For example, to be able to automatically recognise the correct value covered by the individual words considered in their own syntactic context.

Thanks to the LECTIO project, partially supported by the National Agency of I and D (FCT), it has been possible to study new algorithms, in order to better meet didactic needs, and to study new capabilities also useful for research purposes.

One function has been implemented which enables the basic elements of the sentence to be highlighted with different colours. An example of this function is shown in figure 5.

[1] LEMLAT *Analizzatore Morfologico Latino*, is a CNR patent, the authors are: Dr. Andrea Bozzi – Prof. Nino Marinone (for the linguistic aspects) - Dr. Giuseppe Cappelli (for the informatic aspects).

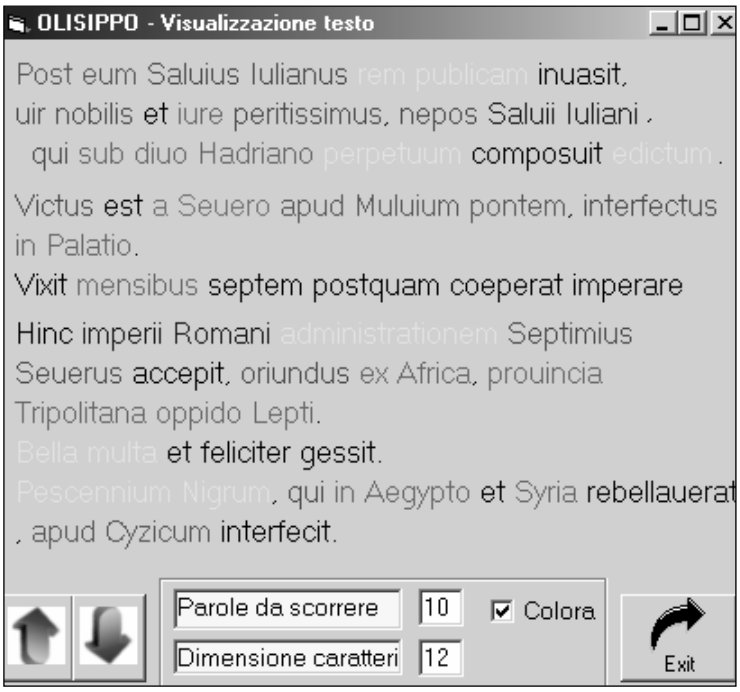

Figura 5

The results of the analysis so far implemented are encouraging, but further work is needed. New algorithms will be implemented very soon, which will improve this module, specially in what concerns the delimitation of sentences.

3. Complementary Remarks

Basically, OLISSIPO is the prototype of a didactic application for acquisition of Latin lexicon Essentially, it produces lists of basic Latin vocabulary enhanced with linguistic information and linguistic extras in an open environment where the user has tools for adapting the prototype according to his specific needs.

If the student is provided with lists of vocabulary taken from texts of these authors (chosen by the teacher according to concrete criteria) his effort for improving his knowledge of new vocabulary will be supported by objective data and the teacher can rationalise and stimulate the effort made. At the same time, the knowledge of vocabulary previously acquired is consolidated.

The statistical results provided by OLISSIPO are of great help. If we want to facilitate, for instance, the study of demonstrative pronouns in the analysis of a text such as Cicero, *Verr*. 4, 48, 106, the diagram which shows the statistical results will support the validity of the chosen text for the lexical and morphosyntactic work.

In order to avoid the possibility of OLISSIPO being a closed tool**,** it has been devised so that it can be reused and associated with other tools in the field of computerised processing of the Latin language (for example, text sources and information sources in general). A considerable part of the methodology developed for OLISSIPO will be reused in the LECTIO project currently being developed at the Centro de Estudos Clássicos da Universidade de Lisboa and the Istituto di Linguistica Computazionale di Pisa (CNR). The new tool, which will be developed other than for didactic purposes, will also be useful for researchers interested in corpus processing.

In brief, OLISSIPO is of great help to students and teachers in their daily work of teaching and learning Latin and also to researchers in the work of linguistic analysis or in the construction of new applications. For the moment**,** it will be available on CD but the possibility of putting it onto the Internet is also being studied.

4. References

- Alberto P., «O projecto Olissipo uma applicação no âmbito do ensino do latim», *Euphrosyne*, 30, 2003, 335-338.
- Bozzi A., Cappelli G., «A project for Latin lexicography: 2. A latin Morphological Analyze», *Computers and Humanities*, 24 V-VI, 1990, 421-426.
- Cappelli G., Passarotti M., «LEMLAT: uno strumento computazionale per l'analisi linguistica del latino – sviluppo e prospettive», *Euphrosyne,* 31, 2003, 519-531
- Delatte L., Evrard E., «S. Dictionnarie Frequentiel et index inverse de la langue latine», Herent, Liegi, 1981.
- Nascimento A., Alberto P., Cappelli G., Pena A., «Identificação automática de elementos básicos da frase latina: o projecto Olissipo», *Euphrosyne* 31, 2003, 515-518.
- Nascimento A., Alberto P., Cappelli G., «OLISSIPO - entre filologia e informática: recursos para gerir o estudo do texto latino», *Euphrosyne*, 32, 2004