
Preface

This special issue is the outcome of the “NLP and Ancient Languages” one-day workshop organized by ATALA in May 2005. At that point we aimed to present the current uses of automated processing techniques in ancient language scholarship. The present collection of articles is intended to highlight the advances of the last four years.

Is there a need to consider “ancient languages” as a subject for NLP? The term first has to be defined, a task not as easy as it might seem. It is easy enough to give examples of the concept, such as Latin, Sumerian, or Ancient Egyptian. The processing of Medieval Arabic or Old French may be added without much hesitation. But where to draw the line? Should we stop at the sixteenth or seventeenth centuries? Should we not rather think in terms of language stages? In short, is there an underlying unity behind the work presented in this volume?

In reality, the best delimitation of ancient languages might be in terms of scholarly community. A number of similar scientific practices, articulated around philology, can be identified. The original text (or often texts), in physical form, is already an object of study. In numerous cases, knowledge of its grammar and lexicon is only partial.

For linguistics in general, the antiquity, and often longevity, of these languages is valuable for diachronic studies on a long-term basis. This last point is explicitly brought up in some of the articles below (see for instance Candido et al.)

For numerous languages, there was until recently no electronic corpus with a claim to being exhaustive. It seems that many projects in recent years have attempted to deal with this problem.

Moreover, the ancient language community is becoming more and more conversant with the tools developed for NLP. The articles in this volume show that recent databases use a rich system of annotations, covering the full spectrum of possibilities. The texts are lemmatized, processed by syntactic parsers, their thematic structure can be tagged, and so on.

A short history of the field

The pioneer here was probably Father Roberto Busa, who as early as 1948-1949 started to compile a corpus of the works of St. Thomas Aquinas, with the help of IBM. This corpus was used for creating concordances and indexes. Like many pioneering projects in the field, this work is still very much alive. Its internet site is now hosted by the University of Navarre.

In the case of Ancient Greek, the first computer-oriented work was carried out in 1956 by Leonard Brandwood in the UK, with a view to finding lexicometric applications. Also in the field of Classics, projects for almost exhaustive corpora were begun in the 1960s. In 1961 in Liège the *Laboratoire d'Analyse Statistique des Langues Anciennes* (LASLA) began to work on Latin literary texts; from 1965 onwards, Greek texts were also involved. Automatic processing was already being used, and lemmatization and morphological tools created. The *Thesaurus Linguae Graecae* (TLG) was developed in 1972 by Theodore G. Brunner. The *Perseus* Project, which originated from this corpus, has been widely distributed, first as a CD-ROM, then as an online database.

In the case of Ancient Egyptian, the earliest work was that of W. Schenkel on the MAAT (*Maschinelle Analyse Altägyptischer Texte*) project in 1967 and the creation of a lemmatized base of some of the coffin texts (which constitute one of the fundamental corpora for this language). In 1984, the workgroup “*Informatique et égyptologie*” produced an encoding system for hieroglyphic texts, based on previous work by Jan Buurman. At that time, however, encoding was still a slow process, and computer databases still rare. Among current projects, we may note the Berlin Academy's *Thesaurus Linguae Aegyptiae*, which brings together a number of more specialized projects for the purpose of building a lexical base for the language. The *Ramsès* project at Liège University is building a large database of Late Egyptian, which should constitute a treebank for grammatical research.

In the 1980s, previous practices were expanded thanks to the advent of the microcomputer, and the idea of computer corpora became more widespread. At that time, the InaLF began what was to become the *Dictionnaire du Moyen Français*, edited by Robert Martin. This project still exists within the wider context of the ATILF (*Analyse et Traitement Informatique de la Langue Française*) laboratory in Nancy.

At that time the actual use of NLP techniques was uncommon, and the focus was mostly on the creation of full-text corpora.

However, the 1990s saw an increasing interest in the structure of such corpora, and especially in the representation of primary sources. Work on manuscripts was developed around the SGML format and the TEI. This work moved naturally onto the new-born World Wide Web. The project initiated in 1990 at Princeton by Alfred Foulet and Karl. D. Uitti on Chrestien de Troyes's novel, *Lancelot ou le chevalier de la charrette*, is a good example of this trend: it was published online as early as 1995.

The articles

A brief comparison of the program of ATALA's 2005 one-day workshop and the articles in this volume shows definite progress.

Mainstream NLP techniques are more and more employed by the teams who work on Ancient Languages. Simple full-text corpora have been replaced (or improved) by lemmatized databases and even treebanks. Perhaps more significantly, their development is guided by more and more sophisticated theoretical questions.

In the first part of this volume, we present four articles on corpus building, all good examples of progress in the field.

The article by Haug et al. deals with a computer corpus of parallel versions of the New Testament in various Indo-European languages. The scientific approach is clearly defined: the goal is to study five types of linguistic phenomena in the languages concerned. The tagging system is quite rich, including such features as the lemmatization of the texts, their syntactic description, and notation of the informational structure. The article includes a careful study of the development of the corpus and a detailed analysis of the possible solutions, and can thus be considered state of the art.

Petrova et al. outline the creation of a corpus of Old High German in order to study the interaction between the enunciative/hierarchical level and syntax from a diachronic perspective. The system architecture and tools used are described in detail, from the text input and linking through the original manuscripts to the search system.

A much later corpus, which extends from the sixteenth to the nineteenth century, is presented by Candido and Aluisio. This is a database of Portuguese texts, built in order to create a historical dictionary. All the steps of the database creation are detailed, with a discussion of existing tools. Two points in particular are developed: lemmatization, which is particularly important in diachronic studies, and the system for editing dictionary entries.

Finally, the article by McGillivray et al. discusses work on the oldest electronic corpus, the Index Thomisticus. The creation of a treebank and a valency dictionary are described. Several approaches to parsing Latin are presented and compared. Once the treebank is built, it is used to enrich the valency lexicon.

Next come a number of works focused on lemmatization and morphological analysis.

Poudat and Longrée examine the behavior of different lemmatization systems applied to Latin. The article is of interest not only for its results, but also for the robustly structured experimental protocol that it exemplifies.

Souvay and Pierrel examine the lemmatization of Middle French, with all the spelling problems involved. They use rewriting rules to represent both morphological phenomena and diachronic variations.

We then switch language family with an article by Barthélémy on the morphological analysis of Akkadian. The formalism he uses, called "multi-grain

morphology," implements multi-band finite state techniques. It is described in detail and compared to other systems for the morphological analysis of Semitic languages.

The article by Kondrak brings us back to general problems of historical linguistics. It describes an algorithm for identifying cognates and phonetic correspondences in lexica, using a combination of three types of information: phonetic similarities, phonological connections, and semantic proximity.

Lastly, the article by Nederhof that ends this issue describes an algorithm for enabling the optimal automatic layout of aligned data, including complex cases.

Thanks

We would like to thank all the reviewers without whom this collection could not have been published: François Barthélémy, Mahé Ben Hamed, Francesco Citti, Gérard Huet, Wojciech Jaworski, Bastien Kindt, George Kiraz, Christiane Marchello-Nizia, Nicolas Mazziotta, Sylvie Mellet, Remo Mugnaioni, Mark-Jan Nederhof, Mark Olsen, Gerald Penn, Sophie Prevost, Wolfgang Schenkel, Richard Sproat, Achim Stein, Paul Tombeur, Laurence Tuerlinckx, Jerzy Tyszkiewicz, and Jean Winand.

Bastien Kindt, Nicolas Mazziotta, Mark-Jan Nederhof, and Sophie Prevost have been particularly helpful, and we would like to extend our special thanks to them here.

Joseph Denooz
Laboratoire d'Analyse Statistique des Langues Anciennes (L.A.S.L.A.),
Université de Liège
Joseph.Denooz@ulg.ac.be

Serge Rosmorduc
Équipe langues et littératures de l'Égypte ancienne,
ÉPHÉ IV^e section/IUT de Montreuil,
Université Paris 8
serge.rosmorduc@genherkhopeshef.org