RAMSES: AN ANNOTATED CORPUS OF LATE EGYPTIAN

J. WINAND & St. POLIS
(F.R.S.-FNRS (Liège))

S. ROSMORDUC
(Paris-VIII)

The *Ramses*-project aims at building an annotated corpus[1] of all Late Egyptian texts[2]. It must be considered as an interdisciplinary project of fundamental research both in Egyptology and in information technology[3]. In Egyptology, *Ramses* should trigger a complete change of paradigm in one's way of studying texts and language, for the encoding will integrate every possible aspect needed for philological and linguistic studies: hieroglyphic script, transcription, translation, morpho-syntactic analysis, semantic and pragmatic features. Those data will constitute the core of the database, but a complete description of the corpus and bibliographical links will systematically be added. As regards information technology, significant progress will be made to help Egyptologists encode texts by using new morpho-syntactic parser automata.

The aim of this paper is to answer four general questions raised by the brief presentation given above: (1) **why** did we consider it necessary to launch such a project? (2) **what** are our goals when speaking of an annotated corpus? (3) **who** will encode the great amount of data needed to make a tool like *Ramses* efficient and who is going to develop the software? and (4) **how** will we manage to achieve our aim?

[1] From a technical point of view, *Ramses* is a relational database in SQL, where the texts themselves are represented and stored in XML. The editing and search software is written in JAVA, and usable both on Mac and PCs. By means of export procedures (XML), the adopted format for the database is fully compatible with what is recommended by the Text Encoding Initiative (http://www.tei-c.org/index.xml).

[2] For the encoding, our objective is to integrate all the material written in Late Egyptian, from the 18th to the 25th dynasty, including the texts written in a softer Late Egyptian ('néo-égyptien mixte' and 'néo-égyptien partiel'), cf. J. WINAND, *Études de néo-égyptien, I. La morphologie verbale*, Ægyptiaca Leodiensia 2 (Liège, 1992), 10-3.

[3] The information technology viewpoint will not be addressed in detail here. Although this project is essentially committed to solving Egyptological problems, it should be stressed that it will also provide a substantial ground for genuine scientific research in information technology: in cooperation with the Department of computer science of the University of Liège, several innovative algorithms will be developed in order to help the input of Late Egyptian texts, and the syntactic analysis thereof.

Why did we launch the *Ramses*-Project?

Despite the high quality of the scientific research done in the study of Egyptian language, significant progress in Egyptian linguistics is hampered by technical problems. To be innovative today, scholars urgently need extensive corpora of texts provided with a complete linguistic analysis; actually, it is our strong belief that research in Egyptian linguistics cannot make significant progress for lack of systematized corpora[4].

Our project is of course not the first one to link Egyptology with information technology; in fact, computers have been part of the Egyptologists' lives for roughly four decades[5]. The foundation of the Round Table "Égyptologie et Informatique" in 1986 was instrumental in this respect[6]. In order to view *Ramses* against the background of the previous achievements, we shall begin with a short reminder, a quick look in the rear view mirror.

First, it is worth mentioning some tools which we are still using today, and which quickly followed the first round table: the "Manuel de codage[7]" (1988) and the very popular software "Glyph for Windows" (1993). Regarding the processing of Egyptian texts, a series of conferences organized in Berlin in connection with the Wörterbuch project deserves special attention[8]. Year 2000 saw the edition of the *Coffin Texts Word*

[4] In this respect, Egyptology lags behind what is being done in Greek and Latin (cf. *e.g.* the *Thesaurus Linguae Graecae* [*TLG*], http://www.tlg.uci.edu/) although databases of classical texts are still far less developed than what could be achieved (see below). Furthermore, one should also keep in mind that Egyptian texts are not available outside the small circle of Egyptologists; this situation should be of concern to all of us from the general perspective of world heritage.

[5] See R. GUNDLACH and W. SCHENKEL, 'M.A.A.T. Ein System zur lexikalischen und grammatischen Erschließung altägyptischer Texte mit Hilfe einer elektronischen Datenverarbeitungsanlage (Projektbeschreibung)', in *Chronique d'Égypte* 83 (1967), 41-64, and the discussions (especially by W. Schenkel, R. Gundlach and J. Leclant) in: A. SCHWAB-SCHLOTT, *Dokumentation ägyptischer Altertümer. Tagung vom 16. bis 17. Juli 1969 in Darmstadt* (Darmstadt, 1970).

[6] Since then, Egyptology in Liège has taken a continuing interest in the applications of information technology to the study of Egyptian, see J. WINAND, 'Analysis of Late Egyptian by Computer', in: G. ENGLUND and P.J. FRANDSEN (eds.), *Crossroad. Chaos or the Beginning of a New Paradigm. Papers from the Conference on Egyptian Grammar, Helsingør 28-30 May 1986* (Copenhague, 1986), 388-400; ID., 'Constitution de fichiers-textes en néo-égyptien: lemmatisation et analyse automatiques', *Revue, Informatique et statistiques dans les sciences humaines* 22 (1986), 179-90; ID., 'Quelques aspects de l'analyse du néo-égyptien par ordinateur', *Informatique et Égyptologie* 4 (1988), 67-80; ID., 'Lemmatisation et levée d'ambiguïté automatiques (II)', *Informatique et Égyptologie* 5 (1988), 76-92; ID., 'Les bases de données de textes en égyptien', *Informatique et Égyptologie* 7 (1990), 161-9.

[7] J. BUURMAN, N. GRIMAL, M. HAINSWORTH and D. VAN DER PLAS, *Manuel de codage des textes hiéroglyphiques en vue de leur saisie sur ordinateur. Manual for the Encoding of Hieroglyphic Texts for Computer-Input. Leitfaden zur Verschlüsselung hieroglyphischer Texte für die Computer-Eingabe*, *Informatique et Égyptologie* 2 (Paris, 1985).

[8] See especially St. GRUNERT & I. HAFEMANN (ed.), *Textcorpus und Wörterbuch. Aspekte zur ägyptischen Lexikographie*, Probleme der Ägyptologie 14 (Leyde, 1999); I. HAFEMANN (ed.), *Wege zu einem digitalen Corpus ägyptischer Texte. Akten der Tagung «Datenbanken im Verbund» (Berlin, 30. September – 2. Oktober 1999)* (Berlin, 2003).

Index[9] and 2004 must be considered as another turning point for it is at that time that the *Thesaurus Linguae Ægyptiae*[10] appeared online. More recently, 2007 can be seen as the *Ramses*-project's official birthdate. Of course, like every child, it was conceived some months earlier[11].

The existing projects illustrate the two major trends in using computers in the field of linguistics for the study of texts[12]:

- lexical databases, as illustrated by van der Plas and Borghouts's *Coffin Texts Word Index*;
- annotated corpuses, the aims and purpose of which are more ambitious. This is the path followed by the *Ägyptisches Wörterbuch*'s team in the Academy of Sciences of Berlin-Brandenburg. This project combines a lexical database with a corpus of texts and has been a valuable source of inspiration for our own project.

Indeed, the *TLA* is the first attempt to design an annotated corpus on a large scale. Its website basically provides two types of information: it first gives access to the original *Zetteln* that were used for the printed version of the *Wörterbuch* and still remain a treasure trove; and it links the user to a new *Textcorpus* whose encoding is still in progress. Such a tool is useful and handy enough for every Egyptologist on a daily basis, but it is fair to say that, at the same time, it falls a bit short of what is expected in linguistics and philology:

- nowadays, the technology at our disposal allows a full integration of the hieroglyphs:[13] the continuing progress made in computer science in the last decades invites us to use without restraint what information technology now has to offer;
- although the texts are lemmatized, the data is not analyzed, neither morphologically nor syntactically. Hence, it turns out to be less helpful for a linguist or a grammarian than for a lexicographer;
- as regards the search option, it is impossible (at least online) to make complex searches by combining two or more lemmas in a query, and there is basically no possibility to sort out the results.

[9] D. VAN DER PLAS and J.F. BORGHOUTS, *Coffin Texts Word Index*, Publications Interuniversitaires de Recherches Égyptologiques Informatisées, 6 (Utrecht-Paris, 1998).

[10] Cf. http://aaew.bbaw.de/tla/

[11] A preliminary presentation of the project is given in St. POLIS, 'Le projet Ramsès', in: J. WINAND, 'Un siècle d'Égyptologie à l'Université de Liège', in: Eug. WARMENBOL (ed.), *La caravane du Caire. L'Égypte sur d'autres rives* (Louvain la Neuve, 2006), 180.

[12] The use of multimedia for the needs of publication and/or diffusion of structured data has not been taken into account here. *Inter alia*, one can mention the online Demotic dictionary made under the auspices of the University of Chicago (http://oi.uchicago.edu/research/pubs/catalog/cdd/), the Deir el-Medina database (http://www.leidenuniv.nl/nino/dmd/dmd.html), or Deir el-Medine online (http://obelix.arf.fak12.uni-muenchen.de/cgi-bin/mmcgi2mmhob/mho-1/hobmain/).

[13] The decision of not encoding the hieroglyphs was taken very early for most probably some good reasons to do so at the time.

The facilities offered by textual databases developed outside Egyptology also appear somewhat limited. For instance, the *Thesaurus Linguae Graecae* (*TLG*) integrates almost every non-documentary text written in Greek between 8th BC and 14th AD. However, as the words are neither lemmatized nor analyzed, the possibilities of research are considerably restricted.

What are our goals?

It was not too difficult to design an ideal structure: we only had to dream eyes wide open. In this perspective, we chose to be very ambitious from the beginning. So we devised a large set of specifications to fulfil scholars' expectations in Egyptian linguistics:

- a hieroglyphic encoding;
- a complete morpho-syntactic analysis;
- a semantic and pragmatic analysis;
- multi-level searches that should:
- be multicriteria, that is, allow not only lexical queries, but also queries on flexions, syntax and semantics;
- be context-sensitive, which means that they should operate at the level of the syntagm, within a proposition, within a sentence, and beyond the sentence (from the paragraph up to the text);
- integrate the graphic level, i.e. hieroglyphs: signs must be searched for by means of a specific coding;
- adapt the corpus of research to specific needs.

And we should be able to do it in some easy way. This means that we never neglected the ergonomics of the software, being always keen on developing user-friendly interfaces. The ambition of the *Ramses*-project is to make encoding as easy as possible and at the same time to ensure coherence by using semi-automatic procedures of analysis.

Who?

The project was launched in 2006 after the Round Table "Informatique et Égyptologie" held in Oxford; it has its roots in a long tradition of interest in Egyptian linguistics and information technology in Liège[14]. We started the project with the human resources of the chair in Egyptology (St. Polis, J. Winand), but quickly brought into the team Serge Rosmorduc, who took part in the database conception, and did the

[14] See n. 6.

programming. In October of the same year, Laurence Neven was hired as research assistant to the project with funding from the University.

In 2007, the programming of several modules had already been achieved and it quickly became evident that the development of the software and the encoding of the data could be done in parallel. Our team grew again in 2008 as we were lucky to get funding from the F.R.S.-FNRS: Stéphanie Gohy and Anne-Claude Honnay (M.A. in Egyptology) then joined the team[15]. And last but not least, the project has been awarded special funding from the University, called an "Action de recherche concertée[16]" (ARC), starting in October 2008; it will allow us to hire over the next five years:

- two young Egyptologists to work on the project and write a PhD on a related topic (cf. n. 15);
- two post-docs in charge of the validation of the encoding and working on the *Prolegomena* for two collective volumes on lexical semantics and on syntactic analysis in Late Egyptian;
- one engineer in computer science to help writing the software and developing the database.

How?

In order to meet all the requirements detailed in our second point, we have designed a database[17] that relies, from the encoder's and the user's viewpoints, on three pillars: (1) a Text-Editor working in close connection with (2) a Lexicon-Editor, and (3) a Search-Engine allowing the user to get every bit of information encoded in *Ramses*.

The Text-Editor

The lemma, the morphology and the exact spelling of each word are encoded in this module (cf. fig. 1)[18]. The morpho-syntactic analysis is done in minute detail; for the first time, multiple analyses for a single unit have been made possible. This provides an elegant solution for treating cases of ambiguity without losing information, which

[15] Each young scholar working on *Ramses* has a PhD subject in close connection with the project: St. Gohy, *Pour une définition du corpus néo-égyptien. Approche linguistique d'une synchronie dynamique*; A.-Cl. Honnay, *Syntaxe générale de la proposition en néo-égyptien*; L. Neven, *Étude du syntagme nominal en néo-égyptien*.

[16] Two main partnerships are to be considered inside the ARC: ÉPHÉ. (S. Rosmorduc, P. Vernus) and LASLA (= Laboratoire d'analyse statistique des langues anciennes, cf. http://www.cipl.ulg.ac.be/lsl.htm).

[17] The encoding of the database is done online; this means that each module has been designed to be multi-user.

[18] For a detailed explanation of the encoding procedure, see A.-Cl. HONNAY and St. POLIS, *Manuel d'encodage du projet Ramsès* (http://www.egypto.ulg.ac.be/Manuel_Ramses.pdf).

means that it is now possible to handle ambiguities in morphology (*e.g.* perfective *sḏm.f vs.* subjunctive *sḏm.f*) and in syntax (*e.g.* sequential *ỉw vs.* circumstantial *ỉw*). This is obviously better than arbitrarily choosing an analysis at the expense of others. Along the same lines, it is also possible to encode a spelling without linking it to a lemma or a flexion in case of doubt about its analysis.

Critical annotations and ecdotic information (lacunae, emendations, palimpsest, dittography, haplography, etc.) are fully taken into account.

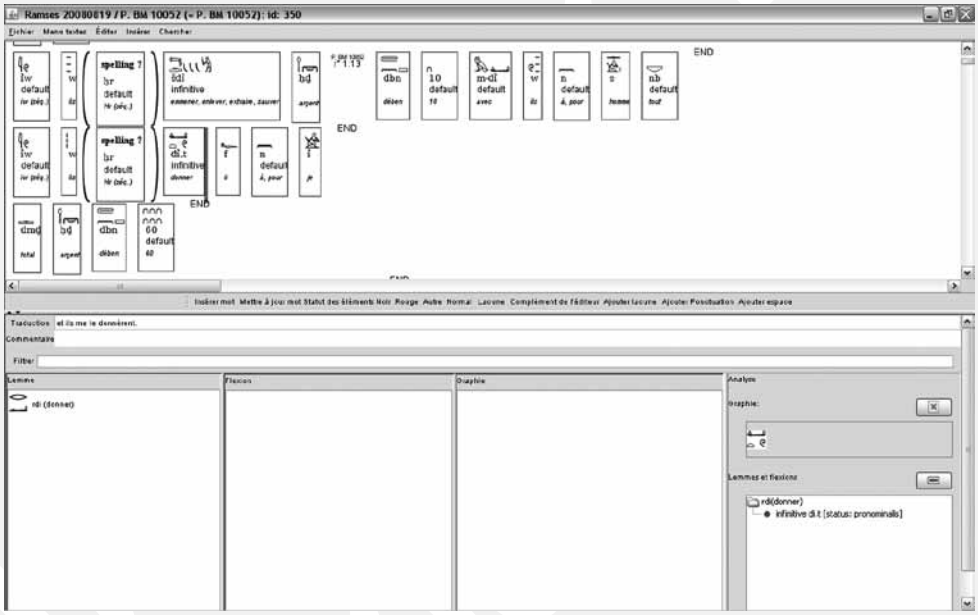

Fig. 1. P. BM 10052, r° 1,12-13 in the Text-Editor.

It must be stressed that the Text-Editor includes a module for describing the corpus of texts: in this module, the type of document, its origin, its date, its medium, its script (hieroglyphs *vs.* hieratic), its level of language and the relevant bibliographical information are recorded. A sharp distinction is made between documents (the actual physical source) and texts; this offers a solution for the encoding of texts recorded in more than one document, like the *Battle of Qadech*.

Lexicon-Editor

Each word encoded in the Text-Editor is linked to a lexical entry (lemma) whose morphological invariants (depending on the part of speech concerned) are accessible

through the Lexicon-Editor. Moreover, this tool enables the encoder to easily add a new lemma, flexion and/or spelling every time a text offers an example of a new form not yet attested in the existing corpus.

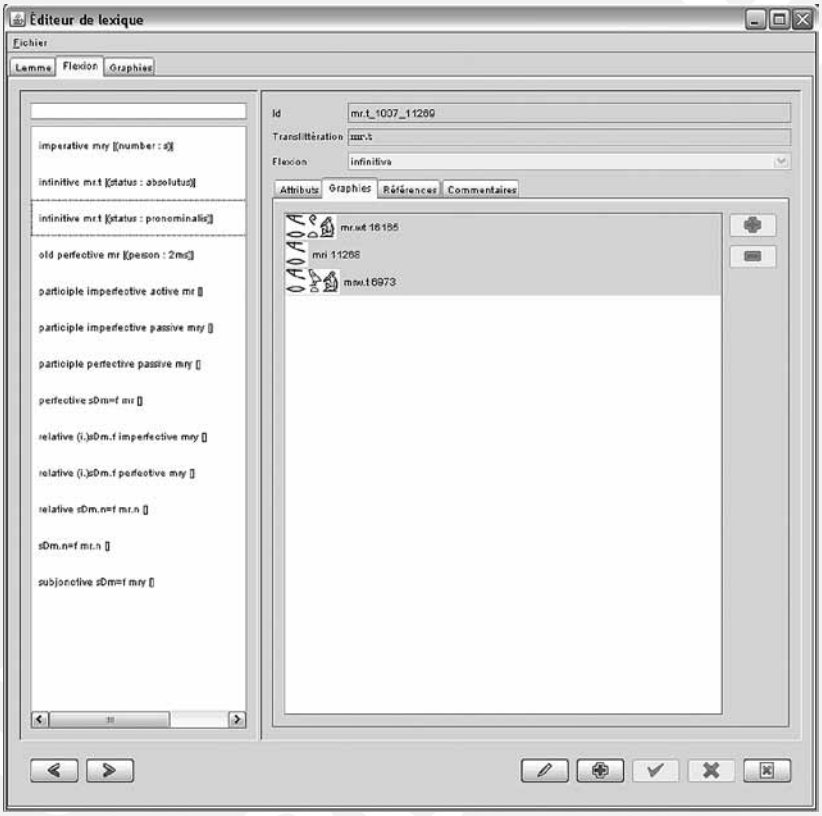

Fig. 2. The Lexicon-Editor: spellings attested for mri (inf. st. pronominalis).

The Search-Engine

We strongly believe that it is by its searching-power that the value of a database can be properly assessed. In this respect, the search facilities offered by *Ramses* go far beyond what exists nowadays in corpus linguistics; it does not seem pretentious to state that *Ramses* allows any kind of research without limitation[19]: it is possible: (1) to build

[19] A more detailed presentation of the Search-Engine is given in: S. ROSMORDUC, St. POLIS and J. WINAND, '*Ramses*. A new Research Tool in Philology and Linguistics', to appear in *Informatique et Égyptologie*.

a corpus of research (using parameters encoded for the texts and/or documents, cf. 4.1); (2) to look for any combination of data (be it lexical, morphological or syntactical, cf. fig. 3); (3) to sort out the results according to the type of data within the scope of the research.

For instance, the figure below illustrates how one can find the collocation of any verb with the walking legs (D54) as determinative followed by a prepositional phrase (r + a geographic name).

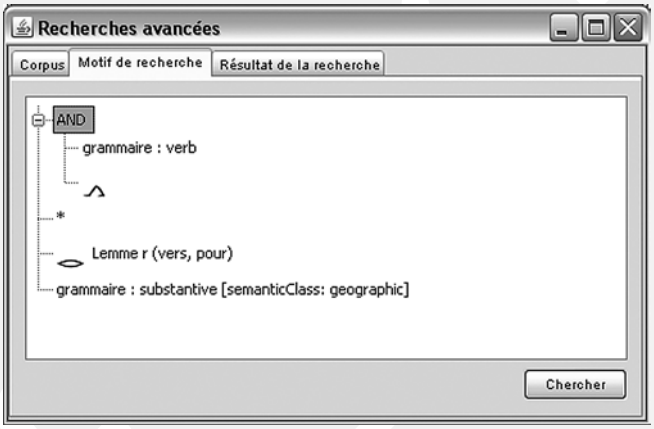

Fig. 3. The Search-Engine.

Conclusions

The modules are at different stages of development, but they are fully operational by now. Currently, there are up to 440 texts already encoded, the number of lemmas in the dictionary roughly amounts to 6,000 and the number of occurrences is about 100,000.

In the near future, the database will be greatly enhanced in at least three ways. First, the number of texts encoded will be dramatically expanded[20]. Second, we will continue developing the software. Many functions still have to be implemented in the Search-Engine (sorting facilities, a better interface for users, graphs and statistics) and the writing of the syntactic and semantic editor will obviously keep us busy for some time to come: this specific tool will be unparalleled in Egyptology and, to the best of our knowledge, there is nothing coming close to it in Classical studies either, as it will

[20] For now, the following sub-corpora are completely encoded in the database: *LES*, *LRL*, *LEM*, *RAD* and the Tomb-Robberies texts.

enable a multi-layered analysis of syntactic features from the word up to the text. Third, a bibliographical module will be interconnected with the documents/texts database, the Lexicon-Editor, and the Text-Editor. So it will be quite easy to add relevant bibliographical notes on general matters (like texts) and on particular points (like words in the lexicon, special spellings, or difficult passages in the texts).

A web interface will enable online searches within the next five years. Once validated at every level, a text will become a fully-fledged member of the *Ramses*-corpus and, hence, available through this user interface.