RAMSES. A NEW RESEARCH TOOL IN PHILOLOGY AND LINGUISTICS

S. Rosmorduc, St. Polis, J. Winand

ABSTRACT

This paper introduces Ramses, a database of Late Egyptian texts, currently under development at the University of Liège (Belgium). Ramses sets out to be a new and powerful research tool. Its main applications are linguistically and philologically orientated. After a general overview of the structure of the database, the search engines are described with some detail.

0. INTRODUCTION

Ramses was officially presented at the Xth International Congress of Egyptologists in Rhodes in May 2008.[1] It is an interdisciplinary project whose purpose is the building of an annotated corpus of all Late Egyptian texts. 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 the 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). We hope that *Ramses* will pave the way for new and innovative approaches to texts and language.

---

1 See J. WINAND, St. POLIS, & S. ROSMORDUC, '*Ramses*. An Annotated Corpus of Late Egyptian', in *Proceedings of the Xth International Congress of Egyptologists*, in course of publication.

The project obtained substantial support from the University of Liège in the form of a so-called *Action de Recherche Concertée* (ARC), starting in October 2008 for a five-year period. In this paper, we examine in some detail (1) the process of text encoding in *Ramses*, (2) the search engine and (3) some prospective developments.

1. ENCODING LATE EGYPTIAN TEXTS IN RAMSES

Any new text must be identified and described. This is done in a dedicated module. The usual information is encoded: writing (hieroglyphic, hieratic), support (ostracon, papyrus, stela, etc.), date, provenance, genre.[2] For the three last features, special lists with a hierarchical structure have been built working more or less like Russian dolls.[3]

When encoding a new text, a clear distinction must be made between documents and texts. As is well-known, there is no necessary overlap between the two: a single text can exist on more than one document, and a single document may host different texts. Information on the writing system, the support and the provenance are stored in the document encoding sheet, whereas information on genre and language are encoded in the text encoding sheet.[4]

Special attention is also paid to the so-called ecdotic information to describe the actual state of a text (lacuna, erasure, words above or under the line, etc.) and the editor's interventions (suppression, addition, restitution, etc.).

One can then proceed to the encoding of the text itself. The text is first segmented in propositions, and propositions are in turn segmented in words. Words are described with three kinds of information: a lemma, a

---

2 The classification of the texts raises many problems. The taxonomy of written Late Egyptian texts is being studied by Stéphanie Gohy. A complete view of the encoding process is given in A.-Cl. HONNAY & St. POLIS, *Manuel d'encodage du projet Ramsès* (http://www.egypto.ulg.ac.be/Manuel_Ramses.pdf).

3 This allows, for instance, to select easily documents by searching the name of a place (*e.g.* the Karnak temple), of an area (*e.g.* the West Bank of Thebes), of a nome (*e.g.* Waset), or of a whole region (Upper Egypt).

4 Illustrations of problems raised by the encoding are given on our web-site (http://www.egypto.ulg.ac.be/Ramses.htm). The dating raises specific problems: in the literary texts, for instance, one has to distinguish clearly between the date of the composition and the date of the copy.

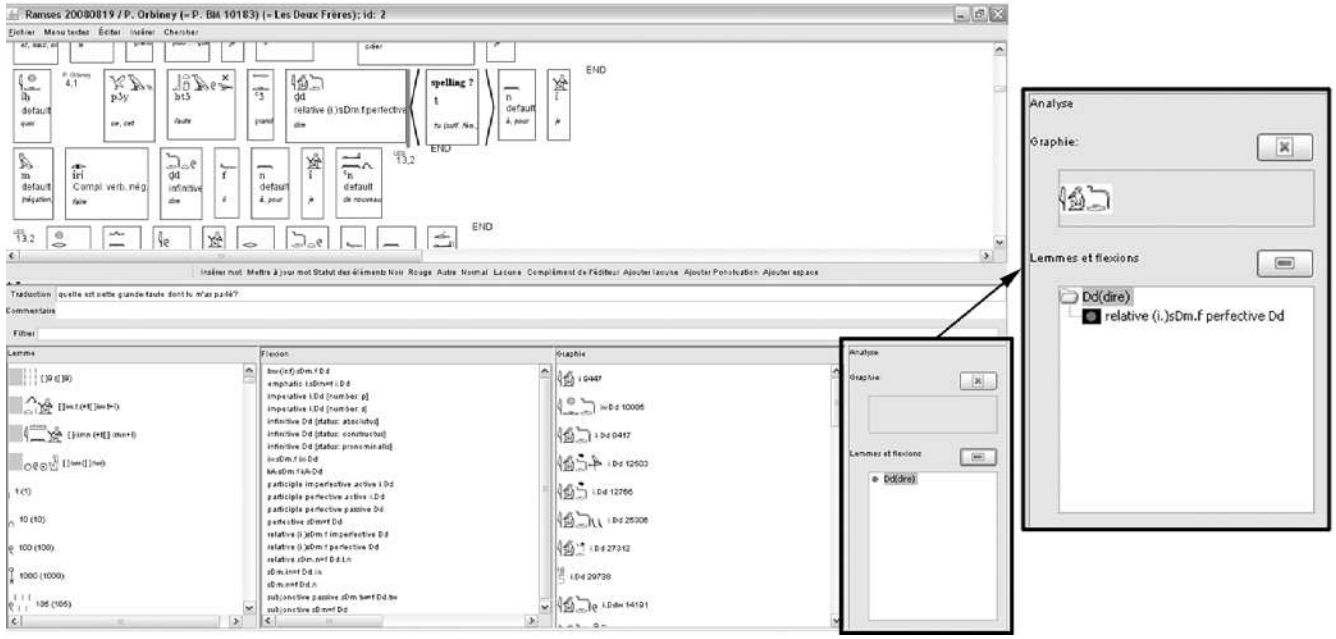

Figure 1. The Text Editor showing an analysis in progress (the final result is displayed in the right box)

spelling, and a morphological analysis (for the syntactic analysis, see below). This information is stored in a related lexicon. By using filters, encoders can quickly select the correct lemma, spelling and grammatical analysis. The lexicon is of course constantly updated as new words, new spellings or new analysis appear.

Figure 2. The Lexicon Editor

Figure 1 shows an extract from the *Tale of the Two Brothers*. The main window, in the upper part of the screen, displays the text. As already noted, the text is segmented in propositions. Words are lemmatized and analyzed, and the spelling is encoded. The lower window gives an idea of the encoding process by showing how the windows look like once the correct lemma has been chosen. This appears in the grey box on the right (see the enlargement). In this example, the cursor stands behind *i·dd*. The second and third columns list respectively all the inflections and spellings already recorded for *dd*.

Using the lexicon, one can immediately see the inflections and spellings that are actually attested for any given word. In the left column of Figure 2 are listed the inflections already attested for this verb in the database; the right column displays the spellings attached to a specific inflections (in this case, the status pronominalis of the infinitive) of the verb *rwi*:

Finally, it was of the utmost importance to find a device to handle elegantly and efficiently ambiguities that can surface at different levels, whether lexical, morphological or syntactical. The grammar of Egyptian can sometimes be a puzzle, making hard to decide between concurrent analyses. Rather than making an arbitrary choice by picking between potentially acceptable analyses, the best solution in those cases is obviously to encode them all. The program has a dedicated routine to treat the ambiguities correctly. This is especially important when figures and statistics are produced.[5]

2. THE SEARCH ENGINE

It is by its power that a database can be gauged, and it is by what it can and cannot do that its value can be properly assessed. If compared with what exists nowadays inside and outside Egyptology, it does not seem pretentious to state that *Ramses* allows (almost) any kind of research without limitation.

The following points deserve some discussion:

- corpus definition;
- search parameters;
- result display.

2.1. Corpus Definition

The search engine offers the possibility to search either the whole database or the text being edited. Now it is also possible to build one's own corpus of research. This can be done by selecting texts on a list, or by using various criteria as filters. For instance, it is possible to restrict the research to the letters written on ostraca during the twentieth dynasty. The results of a previous search can also be used as a corpus for another search, which is often the appropriate way of investigating a problem thoroughly.

2.2. Search Parameters

Basically, any search can involve one or several words. In the latter case, there is the possibility to search adjacent words or to allow some intervening words between. This is achieved by using the skip operator (*). One can search the lemma, the inflection or the spelling, separately or in combina-

---

5 For a concrete example, see our Website (n. 4)

tion. The search engine also allows the use of the boolean operators (AND, NOT, OR).

Here are some examples:

- If one is interested in finding the occurrences of the formula *rdi ḥr.f r/n* X ('to give one's attention to something/somebody'), one must search for the collocations of three distinct words: the verb *rdi* (whatever its inflection), the noun *ḥr*, and the preposition r or n. The subject of the verb does not matter, nor the suffix pronoun after *ḥr*. One must here insert the skip operator to allow any kind of noun phrase. From a logical viewpoint, the request has the following pattern (where a box stands for a word):

Spelling		Spelling		Spelling
Lemma: *rdi*	*	Lemma: *ḥr*	*	Lemma: *n* OR r
Inflection		Inflection		Inflection

In the database, the request will be built as in Figure 3 (using * as a skip operator).

The results are displayed in a list with the text name. One can easily access to the text by double-clicking on the line; the results are highlighted in red (Figure 4).

Figure 3. Request pattern for *rdi Hr.f n/r*

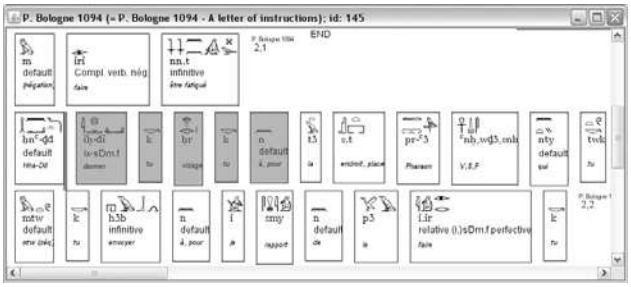

Figure 4. Showing the context of one example of *rdi ḥr.f n/r* (here P. Bologna 1094)

- It is easy to find the occurrences of a inflection without linking it to a lemma. As example, the search pattern in Figure 5 must be used to find the circumstantial perfects (*ỉw sḏm·f*)

Figure 5. The request pattern for the circumstantial perfect *sḏm·f*

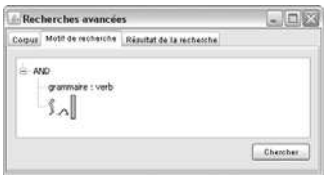

Figure 6. The request pattern for 𓇋𓊃 in verbs

- Spellings are another possible field of research: for instance, one can now search for a string of signs inside a word. This can of course be useful to fill in lacunae, or to study determinatives. In the example below, one looks for the string 𓇋𓊃 in verbs; the query must use the operator AND since two criteria are applied on the same word (Figure 6).

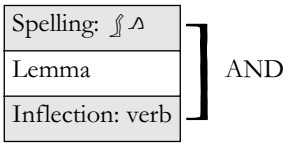

Figure 6. The request pattern for 𓇋𓊃 in verbs

2.3. Display of results

For the present, the program enables only limited facilities for viewing the results. They are displayed in a list that can be sorted out by the documents' name or date. The context can be shown on the screen by clicking on the

corresponding line. These facilities will be greatly expanded in the months to come (see below).

3. FUTURE PLANS

In the next two years, the database should be greatly enhanced in many ways.

3.1. Encoding the Texts

The encoding of the texts started in early 2007. So far, 440 texts have been encoded, which amounts roughly to 100,000 words. There are about 6,000 lemmas in the lexicon.

Ramses aims at encoding all sources written in Late Egyptian. Texts written in mixed Late Egyptian are also taken into account. The time span considered ranges from the 18th to the 25th dynasty.[6] In the coming years, the encoding of the texts will remain a priority. The automata which will be written for handling the syntactic and the morphological issues should greatly help us to reach our goals quickly and efficiently (see below).

3.2. Bibliography

In 2008–2009, a bibliographical module will be written. It will be interconnected with the documents/texts database, the Lexicon Editor, and the Text Editor. So it will be quite easy to add the 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).

3.3. Syntactic Analysis

The next two years will keep us busy with the writing of the syntactic parser. Basically, it will come down to write automata that will be fed with some rules of Late Egyptian syntax. By analysing the context (especially word order), by accessing the classes of words recorded in the lexicon, and by using a statistical approach, the automata should produce reliable hypotheses to group words in syntagms and to assign them a syntactic function.

---

6 J. WINAND, *Études de néo-égyptien, I. La morphologie verbale*, Liège, 1992 (= Ægyptiaca Leodiensia 2).

The advantages of encoding texts with the help of automata[7] cannot be overemphasized: first they really speed the things up, and second – which is undoubtedly as important –, it is probably the best way of ensuring the maximum possible coherence of the data. So, automata could also be written to help the basic encoding of lemmas, spellings and morphological analyses. Needless to say, the final responsibility of the analysis rests upon the human encoder.

3.4. The search engine

The search engine will also be greatly improved. A special focus will be put on the following points:

- extending the possibilities to build a search corpus by making accessible all the descriptors present in the documents/texts database;
- extending the sorting facilities: for the time being, the results can only be sorted according to the date or to the alphabetical order of the document. This should be extended in at least three directions: first, the descriptors being used to build a search corpus should also be available as sorting criteria; second, the lexical and morphological features present in the lexicon should also be used; and third, the syntactic analysis, when completed, should be taken into account. For instance, when looking for all possible occurrences of a circumstantial perfective (see above), the results should be sorted out according first to the verbal lemma, and second according to the spelling;
- the results should be exported in different formats (word processing, spreadsheet, ...) according to one's personal preferences: for instance, taking once again the example of the circumstantial perfective, one should be able to export the occurrences sorted out as suggested above plus, if needed, for each example, the corresponding line in hieroglyphs (with a definable context);

---

7 Multiple approaches are considered here: variations of the very classical *context-free* grammars (É. WEHRLI, *L'analyse syntaxique des langues naturelles*, Paris, Masson, 1997), which are very expressive, but not always convenient for a large corpus, but also grammars based on finite-state automata and transducers, an approach which became popular in Natural Language Processing in the late 1990s (E. ROCHE & Y. SCHABES [ed.], *Finite-State Language Processing*, MIT Press, 1997).

- finally, some statistics should be produced according to the user's choice. This will be particularly welcome when the database will produce vast quantities of results. It will be thus possible to test hypotheses by applying different criteria to see whether they are statistically relevant. For example, when studying the spellings of the personal pronouns, it will be possible to test if the reasons of the variations and changes must be looked for in the diachrony, in the writing system (hieroglyphic vs. hieratic), in the geographical provenance, or in whatsoever still unknown reason.

4. CONCLUSION: WHAT COMES NEXT?

The database will be put on line as soon as possible. Users will first have to register. Basic consultation will be allowed without restriction. The advanced search module will also be accessible for free, but under some conditions.

The database will prove most useful for philological and grammatical studies. In this respect, very innovative topics of research could be addressed especially by young scholars engaged in a doctoral thesis.[8] Researches in the graphic system(s) of Late Egyptian texts are another possibility.

In the long run, we hope to bring to a successful end two main projects which are the most natural outputs of such a database: a dictionary of Late Egyptian, and a complete grammar of Late Egyptian.

---

[8] There are already four PhDs directly related to the *Ramses* project (for more details, see POLIS, ROSMORDUC & WINAND, in *Proceedings of the Xth International Congress of Egyptologists*, n. 15).