The *Ramses* Project
Exploring Ancient Egyptian linguistic data using a richly annotated corpus

Stéphane Polis (F.N.S.-FNRS – ULg) & Jean Winand (ULg)

In this paper, we intend to show how the tools developed for building and exploiting a richly annotated corpus like *Ramses are* changing the methods and practices in Egyptian linguistics. We argue that the wide range of capabilities of Ramses is likely to impact the study of other text languages. Accordingly, a particular emphasis will be put on the new avenues of research that *Ramses* opens up for the study of ancient languages in general.

1. The Ramses project: Introduction

The *Ramses* project — developed at the University of Liège since 2006 (in collaboration with the École Pratique des Hautes Études – Paris) — aims at building a richly annotated historical corpus of the Late Egyptian texts (ca. 1350-700 BCE).

It has been designed with the idea of having a tool specifically dedicated to linguistic research. The corpus includes, for each text, the relevant graphemic (hieroglyphic transcription with transliteration) and linguistic information (complete morpho-syntactic analysis) as well as a full set of meta-data (description and categorization of the corpus, plus bibliographical references). Starting in 2013, we will progressively provide online access to the Ramses corpus.

From a technical point of view, Ramses is a relational database in SQL where the texts are represented and stored in XML. Around 1350 texts have been so far included in the database and received multifaceted annotations. The corpus consisted of slightly more than 300 000 words at the end of 2011 (and is expected to grow up to more than 1 million words in coming years), which amounts to ca. 8000 lemmata, 14 000 inflexions and 45 000 spellings.

2. Structure and capabilities of the Ramses corpus

As a manually annotated corpus, Ramses had to meet two types of basic requirements:

- From the annotator's point of view, the editing software (written in JAVA) had to be user-friendly and to meet the criteria of speed and consistency of annotation.
- From the user's point of view, the annotation schemes should allow for an extreme sensitivity of analysis, but also avoid adherence to any exclusive theoretical linguistic framework, so as to have a wide range of possible end-users.

In order to meet the annotator's requirements in terms of speed and consistency, three interrelated JAVA modules have been designed for handling the graphemic, morphological and syntactic levels: a LexiconEditor, a TextEditor and a SyntaxEditor. We will briefly introduce the principles at work for annotating texts and we will succinctly describe the syntactic formalism, the representational format and the annotation scheme of the SyntaxEditor that aims at building a construction-based treebank of the corpus.

Moreover, we intend to show how we deal simultaneously with three different levels of annotation: [1] corpus mark-up, i.e. meta-data about the texts (textual genre, linguistic register, etc.) and documents (date, nature of the writing support, writing system, place of origin, etc.); [2] ecdotic descriptors (textual criticism is fully integrated); [3] linguistic annotations (that are independent of the graphemic level). Crucially, in order not to freeze the

linguistic information by imposing one particular analysis on a sequence of graphemes, the coding of ambiguities is fully supported by Ramses.

3. Multifaceted annotated corpus and linguistic studies of ancient languages

The latest version of the Search Engine has been the subject of considerable development and now allows (almost) any kind of linguistic query in the corpus. Its nearly unlimited potential will assuredly be of paramount importance for empirically grounded future studies in the fields of graphemics, morpho-syntax, onomastics, lexical semantics, etc. in Late Egyptian.

In this paper we will more specifically address issues related to the study of large-scale phenomena of variation in text languages:

[1] Graphemic variation and argument structure. Ancient Egyptian is a language written in hieroglyphic script, a writing system that uses graphemic classifiers. We show that with the help of Ramses one can — for the first time — study systematically the relationship between the use of graphemic classifiers (in the spellings of verbal lexemes) and argument structure.

[2] Morphological alternation and diaphasic variation. By combining corpus mark up and morphological annotations, we show that some alternations in morphology, far from being free, can be explained by the text genres and registers.

[3] Synchronic variation and language change. Although Late Egyptian is generally considered to be homogeneous during seven centuries of use, Ramses sheds new lights on the gradual emergence of new constructions. We illustrate this point by showing how several parameters interact in the spread of a new vetitive construction.