URN: urn:nbn:de:kobv:b4-opus-24330

STÉPHANE POLIS & JEAN WINAND,
The Ramses project. Methodology and practices in the annotation of Late Egyptian Texts,

in:
Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch“ an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12. – 13. Dezember 2011, herausgegeben von Ingelore Hafemann, Berlin 2013, S. 81-108.

BERLIN-BRANDENBURGISCHE AKADEMIE DER WISSENSCHAFTEN

Thesaurus Linguae Aegyptiae 4

Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch“ an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12. – 13. Dezember 2011

herausgegeben von Ingelore Hafemann

BERLIN-BRANDENBURGISCHE AKADEMIE DER WISSENSCHAFTEN

Thesaurus Linguae Aegyptiae

4

BERLIN 2013

BERLIN-BRANDENBURGISCHE AKADEMIE DER WISSENSCHAFTEN

Perspektiven einer corpusbasierten historischen Linguistik und Philologie

Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch“ an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12. – 13. Dezember 2011

herausgegeben von Ingelore Hafemann

BERLIN

2013

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Dieser Band wurde durch die gemeinsame Wissenschaftskonferenz im Akademienprogramm mit Mitteln des Bundes (Bundesministerium für Bildung und Forschung) und des Landes Berlin (Senatsverwaltung für Wirtschaft, Technologie und Forschung) gefördert

URN: urn:nbn:de:kobv:b4-opus-24310

INHALTSVERZEICHNIS

VORWORT

Die internationale Tagung „Perspektiven einer corpusbasierten historischen Linguistik und Philologie“ vom 12. – 13. Dezember 2011 am Akademienvorhaben „Altägyptisches Wörterbuch“ der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) war dem Thema des Aufbaus und der Nutzungsperspektiven elektronischer Textcorpora und Wörterbücher in den historischen Sprachen gewidmet. Die Teilnehmer, Vertreter der Ägyptologie, der Hethitologie, Indogermanistik sowie Referenten aus der historischen Lexikographie des Mittel- und Frühneuhochdeutschen und des Altfranzösischen diskutierten vor allem über die Veränderungen, die mit dem Einsatz elektronischer Erfassungs- und Verarbeitungsprozeduren einhergehen. Vertreter der Computerlinguistik vom „Zentrum Sprache“ der BBAW wurden in die Diskussionen einbezogen. Dort beschäftigt man sich seit Jahren mit dem Aufbau großer elektronischer Textcorpora (DWDS), darunter auch solcher, die historische Texte (DTA) für die elektronische Nutzung ermöglichen.

Die größte Herausforderung dieser neuen elektronischen Corpora und Wörterbücher ist es, sowohl den Methoden und damit den wissenschaftlichen Ansprüchen der traditionellen Philologie und Lexikographie unbedingt verpflichtet zu bleiben als auch neue Gebiete wie die Corpus- und Computerlinguistik für die historischen Sprachen zu öffnen. Die Teilnehmer haben gemeinsam und disziplinenübergreifend die Möglichkeiten und Grenzen der Datenerfassung, ihrer Präsentation und den Nutzen neuer Auswertungsprozeduren diskutiert.

Unter dem ersten Thema „Historische Corpusprojekte – synchron und diachron“ wurden elektronische Corpora vorgestellt und ein intensiver Austausch darüber geführt, welche Datenstrukturen die linguistischen Inhalte in adäquater Weise abbilden. Wichtig war die Frage, auf welche Resonanz diese elektronischen Corpora bei den Nutzern gestoßen sind und welche Erwartungen und Anforderungen aus den verschiedenen Fachdisziplinen an die Projekte herangetragen werden. Der Austausch über Nutzungsperspektiven elektronischer Corpora schloss auch die Diskussion über die Erarbeitung projektübergreifend einsetzbarer Standards der Codierung und Strukturierung historischer Textdaten mit ein. Hinsichtlich einer mittel- und langfristigen Nutzbarkeit sowie einer langfristigen Datensicherheit stehen solche Fragen zunehmend im Focus und einige aktuelle Initiativen dazu wurden vorgestellt. Spezielle technische Aspekte

elektronischer Datenerfassung und automatischer Analyse- und Speicherungsverfahren elektronischer Textdaten konnten am letzten Tag als ein Themenschwerpunkt mit den Programmierern diskutiert werden.

Ein zweiter Schwerpunkt waren konkrete Fragstellungen aus der historischen Lexikographie und diachronen Textanalyse. Für das Ägyptische ist der diachrone Ansatz auf Grund der über viertausendjährigen Textüberlieferung von großer Relevanz. Themen wie historischer und/oder textgattungsspezifischer Wortgebrauch, die Erarbeitung diachroner Wortlisten und Aspekte des kontaktindizierten Sprachwandels konnten disziplinübergreifend zwischen den Ägyptologen und den Kollegen der historischen Lexikographie des Mittel- und Frühneuhochdeutschen und des Altfranzösischen behandelt werden.

Mit dem Abendreferenten Gregory Crane, dem Begründer der „Perseus Digital Library“, wurde ein breites Publikum angesprochen. In seinem Vortrag hat er noch einmal die hohe Relevanz und die neuen Möglichkeiten der Einbeziehung zahlreicher Wissenschaftler und einer interessierten Öffentlichkeit in die Projektarbeit demonstriert, die das Internet auf völlig neue Weise eröffnet hat. Die Herausgeberin ist sehr froh, seinen programmatischen Beitrag zu diesem Thema, dessen schriftliche Form er gemeinsam mit Alison Babeu erarbeitet hat, ebenfalls in diesem Band präsentieren zu können.

Wir danken der Berlin-Brandenburgischen Akademie der Wissenschaften für die umfassende Unterstützung unserer Projektarbeit und ganz speziell der Vorbereitung dieser Konferenz sowie der Möglichkeit, die Akten auf dem E-Doc-Server der Akademie veröffentlichen zu können.

Der Hermann und Elise geborene Heckmann Wentzel-Stiftung sei hiermit ausdrücklich für die unbürokratische und großzügige finanzielle Unterstützung dieser erfolgreichen Tagung gedankt.

Das Akademienvorhaben „Altägyptisches Wörterbuch“ konnte sich als aktives Mitglied des Weiteren auf das „Zentrum Grundlagenforschung Alte Welt“ stützen, dem alle altertumswissenschaftlichen Vorhaben der BBAW angehören. Dem Zentrum ist es zu danken, dass der Abendvortrag von Gregory Crane einem breiteren Publikum dargeboten werden konnte.

Allen Autoren dankt die Herausgeberin für ihre anregenden Diskussionen und die qualitätvollen Beiträge in diesem Band.

Auf eine Gesamtbibliographie wurde verzichtet und die Abkürzungen der in den ägyptologischen Beiträgen erwähnten Zeitschriften und Reihen folgen dem Lexikon der Ägyptologie, herausgegeben von Wolfgang Helck und Wolfhart Westendorf, Band VII: Nachträge, Korrekturen, Indices, Wiesbaden 1992, XIV-XIX.

Ganz besonders sei schließlich Frau Angela Böhme für die gewissenhafte redaktionelle Bearbeitung der Manuskripte gedankt sowie Dr. Simon Schweitzer für seine Hilfe beim Erstellen des Layouts.

Berlin, Mai 2013 Ingelore Hafemann

THE RAMSES PROJECT
METHODOLOGY AND PRACTICES IN THE ANNOTATION OF LATE EGYPTIAN TEXTS

STÉPHANE POLIS & JEAN WINAND

0. Introduction

This paper is an updated presentation of the Ramses project being currently developed at the University of Liège.[1] The first section stresses the main objectives and gives a technical description of the general architecture of Ramses software.[2] The second part describes the encoding procedures and reviews the current state of the annotation. In the third section, some changes brought about by the use of large-scale corpora are discussed from an epistemological viewpoint. The paper ends with the presentation of some new avenues for research that will ensue from the use of a complex multilevel corpus.

1. Goals and Means

1.1 The philosophy behind Ramses

The Ramses project that has been under development in Liège since the end of 2006 is deeply rooted in the fields of expertise of its creators. This explains some critical decisions that have been made

---

[1] Previous reports are POLIS, S., 2006: Le projet Ramsès, in: WINAND, J., Un siècle d'Égyptologie à l'Université de Liège, in: WARMENBOL, E. (ed.), *La caravane du Caire. L'Égypte sur d'autres rives*, Louvain-la-Neuve, 180; ROSMORDUC, S. *et al.*, 2009: Ramses. A new research tool in philology and linguistics, in: STRUDWICK, N. (ed.), *Information Technology and Egyptology in 2008. Proceedings of the meeting of the Computer Working Group of the International Association of Egyptologists (Informatique et Égyptologie), Vienna, 8-11 July 2008*, Bible in Technology 2, New Jersey, 133-142; POLIS, S. *et al.*, 2013: Building an annotated corpus of Late Egyptian. The Ramses Project: Review and perspectives, in: POLIS, S. & J. WINAND (eds.), *Texts, Languages & Information Technology in Egyptology. Selected papers from the meeting of the Computer Working Group of the International Association of Egyptologists (Informatique & Égyptologie), Liège, 6-8 July 2010*, Ægyptiaca Leodiensia 9, Liège, 25-44; WINAND, J. *et al.*, Forthcoming: Ramses. An annotated corpus of Late Egyptian, in: KOUSOULIS, P. & N. LAZARIDIS (eds.), *Proceedings of the Tenth International Congress of Egyptologists. University of the Aegean, Rhodes, 22-29 May 2008*, Orientalia Lovaniensia Analecta, Leuven, 10 p.

[2] From a technical point of view, Ramses is a relational database in SQL where the texts are represented and stored in XML; the software interface is written in JAVA.

from its very inception. Ramses is both a philological and a linguistic tool, with perhaps more emphasis on the latter dimension. The database is intended to answer all possible questions that can arise when studying a text language. Such a goal is admittedly very ambitious — and might even sound pretentious —, but given the present technical means, it seems far from being unrealistic.

Indeed, most databases presently available — for ancient text languages and for modern languages alike — are usually very good at retrieving isolated words, with varying degrees of precision when it comes to grammatical inflexions. However, they perform less efficiently when it comes to complex queries concerned simultaneously with several layers of annotation. The situation can even become inextricable if these layers are combined within queries that involve several words, phrases or sentences. It is those kinds of shortcomings that the general architecture of Ramses will hopefully overcome.

Moreover, Ramses has been developed with an evolutionary database design: it has the capability of integrating new layers of annotation (that will eventually be connected to the pre-existing levels of annotation). For instance, it would be possible to add a new layer of analysis for tagging proper names with all relevant socio-professional information and to use it as a filter when analyzing the textual data. This, of course, would be a major improvement for those interested in studying prosopography in relation to written production.

1.2 Software Architecture: The relationship between the modules

As a richly annotated corpus, Ramses required software capable of a fair degree of complexity. The first figure (Fig. 1) gives a schematic overview of the general architecture of Ramses' software.

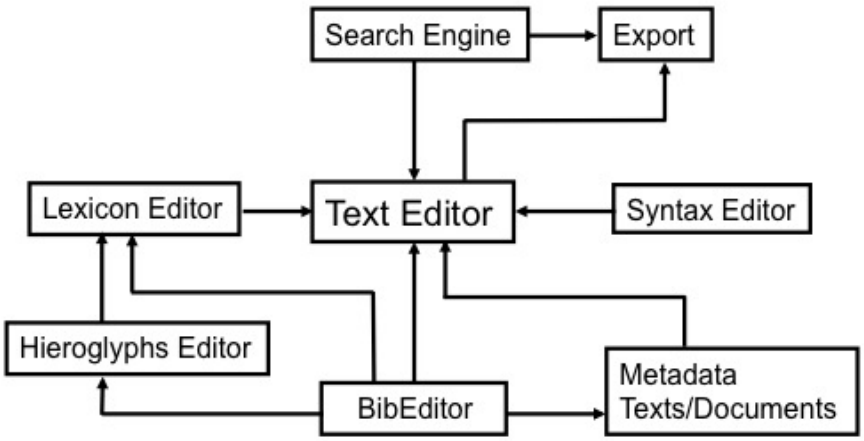


Figure 1. Software architecture of Ramses

1.2.1 The annotation tools: Lexicon, morphology and syntax

The **TextEditor** is the core module. This is the part of the interface that first presents itself on the screen when the database is opened by one of the annotators.[3] The text is segmented in words (Fig. 2); each word is graphically isolated in a box that contains some basic information (Fig. 2, box 1):

- The hieroglyphic spelling;
- The transliteration and the label of the inflexion;
- The standard translation of the lemma.

[3] The interface will obviously be adapted for end-users when we make Ramses available online.

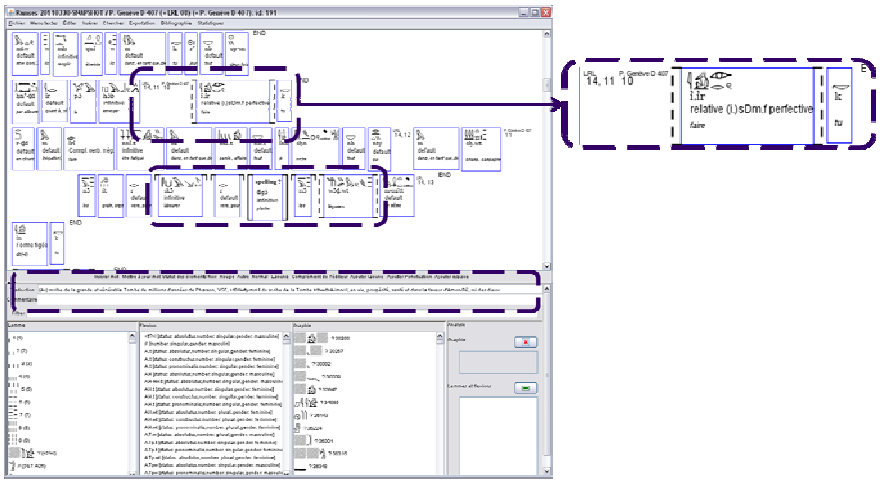


Figure 2. TextEditor

Textual criticism is entirely integrated: lacuna, editor's restoration, erasure, etc. are systematically annotated (Fig. 2, box 2).

For each text, there is a double reference system: first according to the document's materiality (e.g. r° 1,2), second following the modern edition that has been used; a marked preference has been given to well-known collections of texts like *LRL*, *LES*, *KRI*, *LEM*, etc. A translation in French or sometimes in English (depending on the annotator's first language) is provided at the bottom of the main window; it is aligned sentence by sentence (Fig. 2, box 3).

The three lists at the bottom of the screen (see Fig. 2) contain all the lexemes, inflexions and spellings already recorded in the database. Thanks to basic statistical functions, filters help the encoders to find the adequate analysis in context when annotating new occurrences. The result appears in the last box on the right.

Those lists are connected to the data encoded in the **LexiconEditor**. Fig. 3 illustrates what is displayed in this module when the verb *irỉ* "to do" has been selected. Within the central window, all the spellings that have been encoded so far for this verb are displayed: 257 different spellings are stored so far in the database for this very common verb.

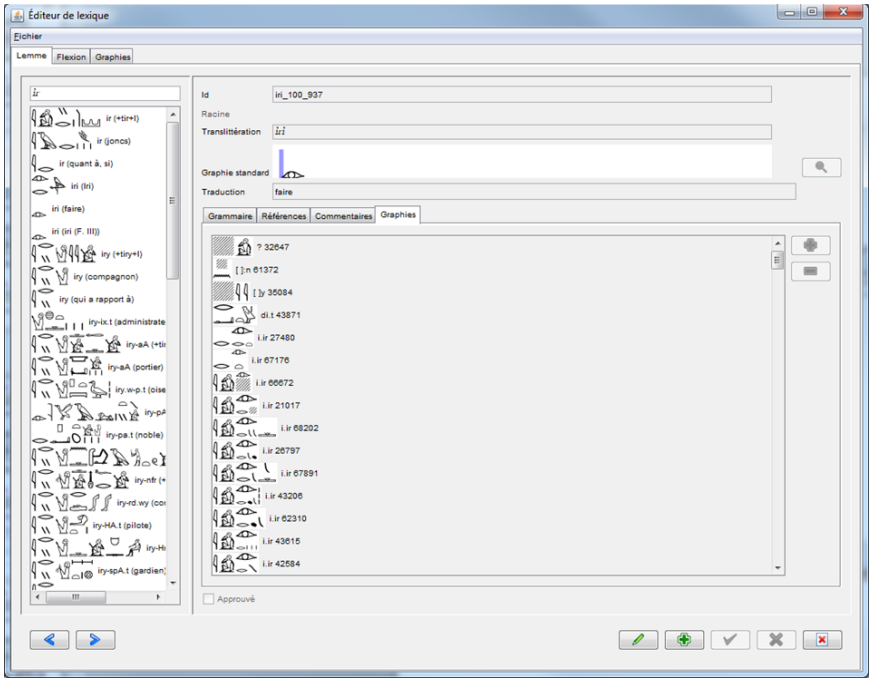


Figure 3. LexiconEditor

Once a lemma has been selected, it is possible to visualize the inflexions that have been linked to it (Fig. 4). If the user picks one of them (e.g. the emphatic form *i.ir=f*), the spellings that have been annotated for this particular inflexion appear in the main window. It is worth noticing that one can also proceed the other way around — which can prove to be extremely useful: starting from a particular spelling, it is possible to visualize all the inflexions that have been linked to it in the corpus.

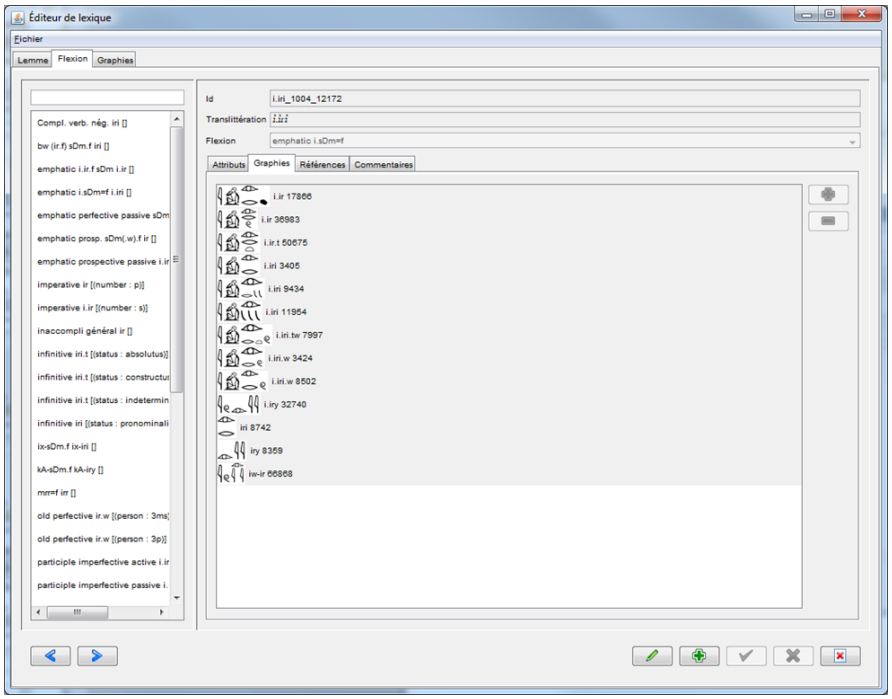


Figure 4. LexiconEditor – Spellings of the "Emphatic (i.)sḏm=f" inflexion

In order to create a new hieroglyphic spelling, a special module has been designed, the **HieroEditor** (Fig. 5), an offspring of Serge Rosmorduc's JSESH hieroglyph editor,[4] that basically works along the principles of the *Manuel de Codage* (with slight modifications and additions).

[4] See http://jsesh.qenherkhopeshef.org.

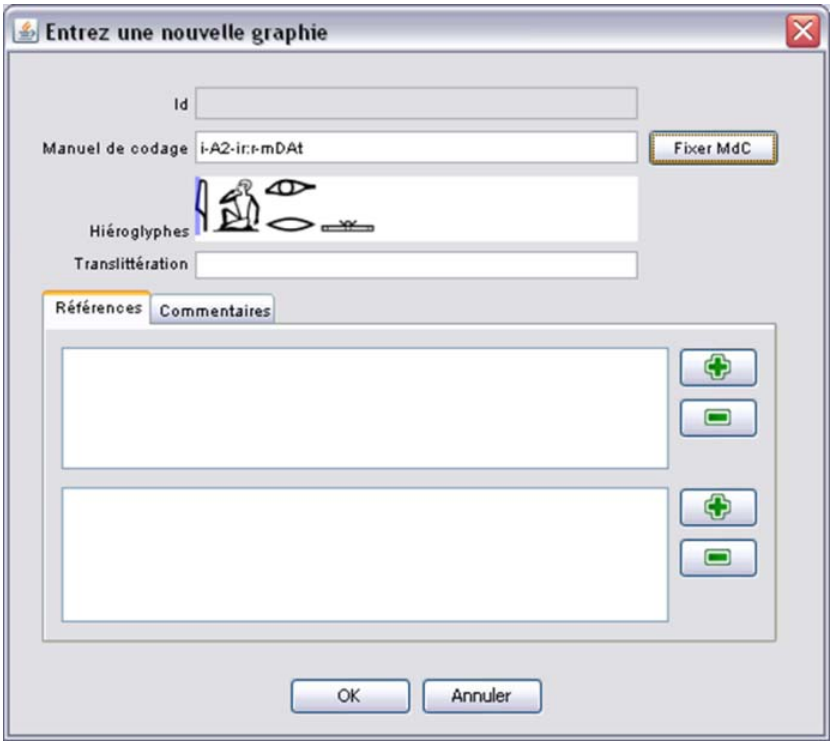


Figure 5. HieroEditor

The **SyntaxEditor** is still under development, but is already in a test-phase.[5] It capitalizes on the data annotated in the TextEditor, and makes them fully available when one performs syntactic annotation.

The functionalities of the SyntaxEditor have been developed in order to allow not only phrasal chunking (supporting discontinuous constituents, as in the simplified Ex. of Fig. 6) and full syntactic analysis of a sentence, but also in order to annotate other dimensions of linguistic analysis like anaphoric relations (field of textual cohesion,

[5] For a complete description of the SyntaxEditor (specifications that are implemented, syntactic formalism, representation format and annotation scheme), see POLIS, S. & S. ROSMORDUC, 2013: Building a construction-based treebank of Late Egyptian: The syntactic layer in Ramsès, in: POLIS, S. & J. WINAND (eds.), *Texts, Languages & Information Technology in Egyptology. Selected papers from the meeting of the Computer Working Group of the International Association of Egyptologists (Informatique & Égyptologie), Liège, 6-8 July 2010*, Ægyptiaca Leodiensia 9, Liège, 45-59.

e.g. via the co-indexation of pronouns and noun phrases) and information structure as well as speech acts.

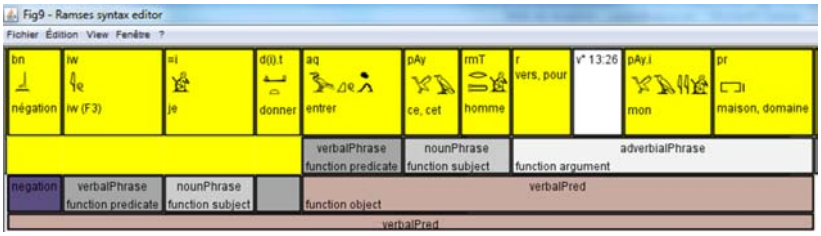


Figure 6. SyntaxEditor

The annotation scheme, which defines the valid types of syntactic annotations as well as the possible set of functions, construction by construction, is a priori neither framed in a constituent structure nor in a dependency-based formalism: we see these representations as two different possible outputs of a single 'construction-based' annotation scheme. This approach — close to the one developed in Potsdam university for the TIGER Treebank[6] — has been developed in order to account for the diversity of linguistic facts found in the Late Egyptian corpus. It is much in agreement with the grammatical tradition in Egyptology, which endorsed a *construction grammar* perspective *avant la lettre* by systematically taking into consideration different grammatical *patterns*.

This perspective takes seriously the assumption that *constructions* are the basic units of any syntactic representation. Accordingly, we consider as a real possibility that the syntactic annotation will lead to generalizations concerning elements across constructions that are not congruent with the pre-existing categorization (e.g. parts-of-speech that are encoded for each lemma in the LexiconEditor). This means that the syntactic annotation will most certainly have feed-back effects on the previous analyses, thereby avoiding the methodologically untenable position of defining a priori categories such as part-of-speech, etc.

From an IT point of view, the TextEditor and the SyntaxEditor will eventually merge into a single JAVA module with visualization facil-

[6] BRANTS, S. *et al.*, 2002: The TIGER Treebank, in: *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, 24-41.

ities that will enable the annotators to select the level of linguistic analysis they wish to have access to.[7]

1.2.2 Appending metadata to the corpus

The annotation of the linguistic material would be virtually useless without metadata. These are recorded in Ramses with the help of two main modules: the **TextDocumentEditor** and the **BibEditor**.

The annotated texts are identified and described in the **TextDocumentEditor** (Fig. 7). Texts and documents as material objects must be carefully distinguished.[8] In most of the cases, the two categories overlap, but a text is sometimes preserved on many documents (this is of course mostly the case with literary and religious texts; all the parallel versions of a given text are annotated — and will be later aligned — in Ramses), and a single document can also contain more than one text, as is the case with anthologies, for instance.

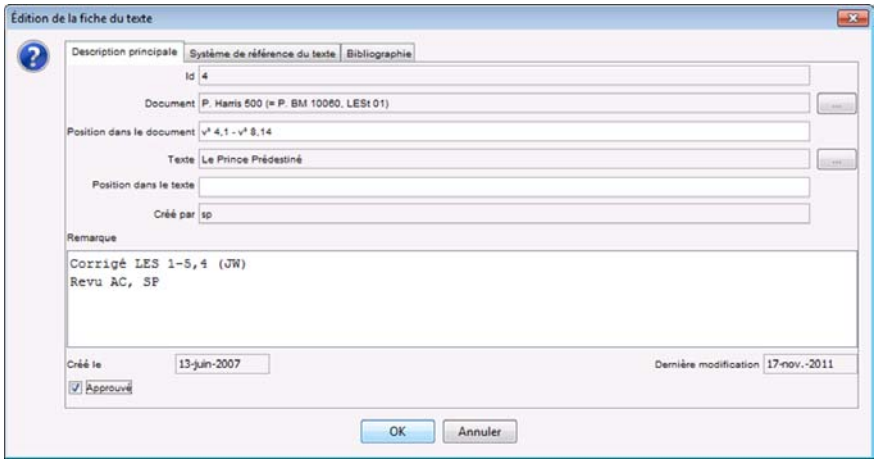


Figure 7. TextDocumentEditor

[7] The SearchEngine for the syntactic layer is not implemented yet. We currently investigate the possibility of using *Annis2* (see http://www.sfb632.uni-potsdam.de/annis/), "an open source, versatile web browser-based search and visualization architecture for complex multilevel corpora with diverse types of annotation."

[8] Cf. the distinction between object and text in the *Thesaurus Linguae Aegyptiae*.

Metadata such as date, provenance, writing system, writing support, language sub-categorization, textual genre, are based on hierarchical thesauri that match recognized standards such as the *Multilingual Egyptological Thesaurus*[9] whenever possible.

Furthermore, modern literature can be appended selectively to different levels of annotation (see Fig. 8) in order to justify the choices and interpretations made by annotators.

Complete references are first encoded in a specialized **BibEditor**. They are then linked, with the appropriate pagination and tags specifying their content, to different objects of the database. The following screen shot (Fig. 8) shows how bibliographical references are instantiated in the LexiconEditor for the lemma *ỉb* "heart" (especially noteworthy are the hyperlinks to other electronic resources such as the *Thesaurus Linguae Aegyptiae*[10], *The Deir el-Medina Database*[11], *Deir el Medine Online*[12] and the *Online Egyptological Bibliography*[13]).

[9] See VAN DER PLAS, D. (ed.), 1996: *Multilingual Egyptological Thesaurus*, Publications Interuniversitaires de Recherches Égyptologiques Informatisées 11, Utrecht/Paris.

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

[11] See http://www.leidenuniv.nl/nino/dmd/dmd.html.

[12] See http://dem-online.gwi.uni-muenchen.de/.

[13] See http://oeb.griffith.ox.ac.uk/.

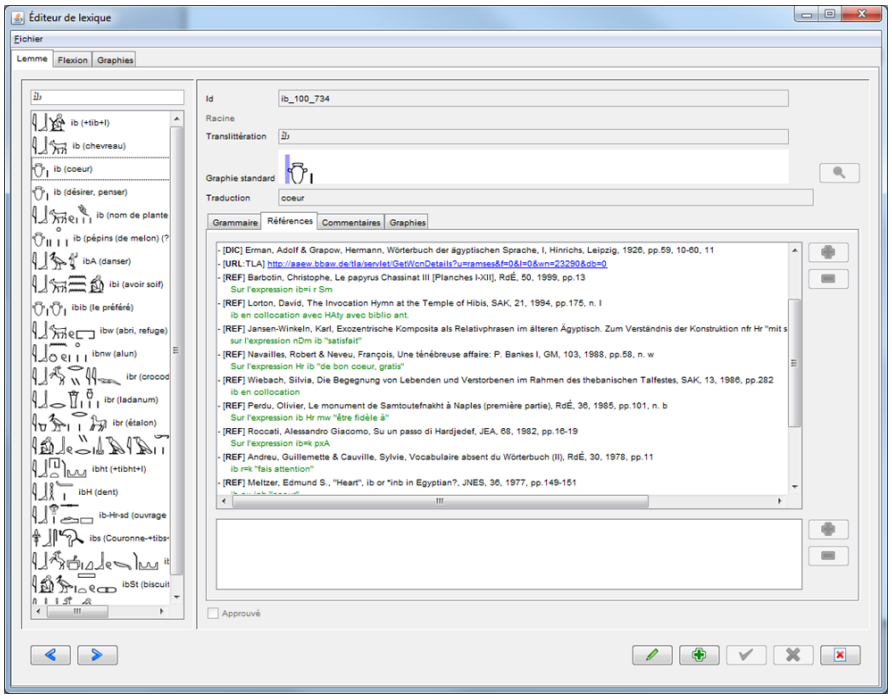


Figure 8. BibEditor

1.2.3 The SearchEngine

A database, however rich and complete, is useless without a powerful system for retrieving the relevant information. As noted above (see §1.1), the SearchEngine has been designed to run ideally any type of queries, without limitation regarding the types of annotations or metadata that can be searched for simultaneously.

Queries can be made on the whole corpus or on sub-corpora by using filters on genres, date, provenance, writing support, writing system, and so on. Fig. 9 shows how one can build a query on a sub-corpus containing all the letters that have been written on ostraca and come from the village of Deir el-Medina.

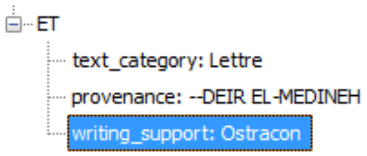


Figure 9. Defining a subcorpus

Any query is built step by step. One specifies successively the layers of annotation that are searched for and the context that will be taken into consideration. In the following example (Fig. 10), the search aims at finding fronted relative clauses that are introduced by *ir* and whose predicate is a verb that has both the infinitive as inflexion and the moving-legs as classifier. The skip operator (*) that appears twice in this example means that unspecified words are allowed between two elements of the query. If needed, the number of these un-accounted blocks can be more or less strictly specified (exactly 3 occurrences or between 1 and 4 occurrences).

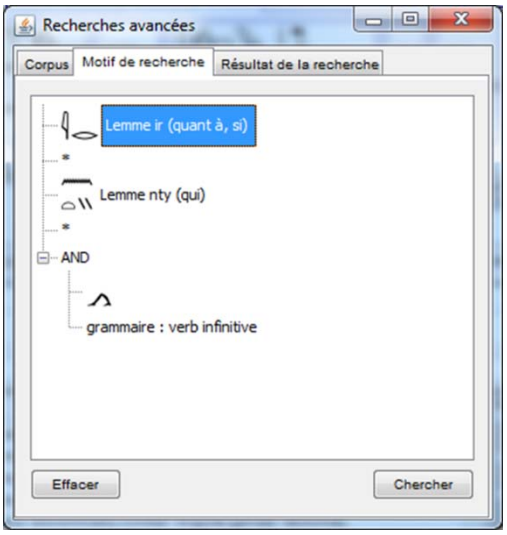


Figure 10. SearchEngine

The results can of course be visualized in a table format. Each line of the results is linked to the TextEditor, so that the end-user can easily access a wider and fuller context with the relevant bibliography, if any.

Finally, the data can be exported in `.pdf`, `.html` or `.gly` file format. All levels of annotation can be exported at once, but it is also possible to select specific data to be exported. In the first example (Fig. 11a), all the data have been exported in `.pdf` format; it should be noted that interlinear grammatical glosses are produced automatically, based on the annotated data. The second example (Fig. 11b) illustrates a lighter option: the hieroglyphic line has been exported in `.gly` format,[14] without the lexical and grammatical tagging.[15]

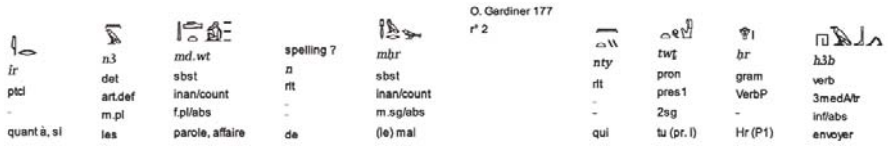


Figure 11a. Export Tool (a sentence in `.pdf` format)

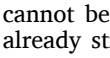
O. Ashmolean Museum 0177 (= O. Gardiner 0177); id: 1404

Figure 11b. Export Tools (same sentence in `.gly` format)

2. Building an annotated corpus: Methodology and current state

In this section, the current state of the annotation process is reviewed with a particular emphasis on the way an annotated corpus like Ramses is actually built. In the first part, we focus on the methodological principles at stake when annotating texts in the corpus and we show how software developments have been used in the fight against time, probably enemy number one in the lengthy task of

[14] The `.gly` export format has recently been implemented by Serge Rosmorduc. It proves to be an especially useful and time-saving tool when data coming from Ramses are used in a later written production.

[15] As the hieroglyphic line is composed automatically by juxtaposing the coding of the individual blocks, the relative position of the signs at the border of two words cannot be accounted for. A sequence like  will appear as . As already stressed, Ramses is primarily a research tool with a clear orientation to questions related to grammar and linguistics; it will never substitute for a sound philological edition nor for a photograph.

building a corpus. In the second part, we comment upon figures summarizing the progress made so far in the encoding and in the annotation of the textual data. Finally, future prospects are outlined in the third part.

2.1 Software ergonomics

As a manually annotated corpus, Ramses had to meet one requirement of paramount importance from the annotator's point of view: the editing software had to be user-friendly so as to meet the criteria of speed (and ideally consistency) of annotation.

In order to meet this requirement, three interrelated JAVA modules (see §1.2.1) have been designed for handling the graphemic, morphological, syntactic and textual levels: a TextEditor, a LexiconEditor and a SyntaxEditor. We will focus here on the relationship between the first two modules when annotating a text.

The goal was to save annotators from reduplicating work by implementing fully the capabilities of relational databases. Therefore, the following principle has been adopted: each occurrence of a word in a text (TextEditor level) is the actuation of a detailed entry in the lexicon (LexiconEditor level).

In Fig. 12, for instance, the verb *gmḥ* in one sentence from the *Doomed Prince*, is an actuation of the lemma *gmḥ* in the LexiconEditor (on the left) and of the inflexion /infinitive_StatusConstructus/ (on the right).

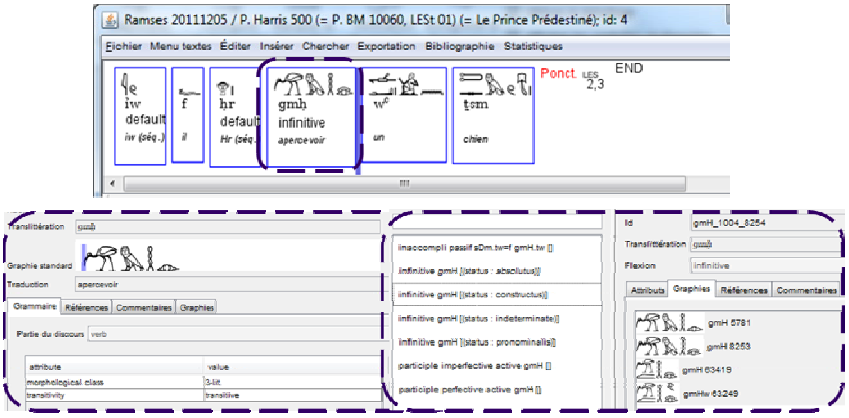


Figure 12. Link between the TextEditor and the LexiconEditor

When encoding a text in the TextEditor, the annotator simply has to select the lemma, the inflexion and the spelling from lists (see Fig. 2) that are fed by the LexiconEditor and sorted according to basic statistics automatically generated about the existing corpus.[16] If any lemma, inflection or spelling is missing, these lists are supplemented by adding new information in the LexiconEditor.

The encoding of texts was obviously quite slow at the beginning of the project (given that every single new occurrence had to be fully encoded in the LexiconEditor), but as the corpus was growing and the data in the LexiconEditor correlatively expanding, the annotator's work became correlatively faster: annotators never have to encode the same data twice. At every single level — from inflexions to spellings, from bibliographical references to documents and texts — data are encoded only once and they are directly available and easily accessible for the all the annotators working on the database.

2.2 Progress in the annotation

Whatever the quality of the tools developed for facilitating the encoding, Ramses remains a manually annotated corpus, which means that the integration of new texts in the database is time consuming.

Besides software developments, an additional strategy has been devised in order to speed up the process of annotation (and hopefully to increase its consistency): since Late Egyptian written registers are highly diverse — in terms of lexicon, phraseology, distribution of inflectional patterns, etc. — the whole Late Egyptian corpus has been split up into sub-corpora according to text genres and chronological periods. Each annotator working in the project is responsible for the annotation of a particular *Textsorte*.

Currently,[17] 1744 texts have been worked out in the database and received multifaceted annotations, which amounts to a little more than 334,000 tokens or words. As shown by Fig. 13a-b, the progress made in the encoding is quite regular. The last two years even testify a slight increase of the number of new words annually annotated in

[16] We are currently developing a context-sensitive semi-automatic tagger that suggests the lemma, inflection and spellings that are the most likely to be accurate for a word (taking into account mark-up data such as the genre, date and support of any new text). This tool should significantly enhance the speed of annotation.

[17] The statistics provided below have been produced on 2012/09/15, which means that the figures for 2012 are not complete yet.

the database, which has resulted from capitalization on the strong base of a well-stocked LexiconEditor.

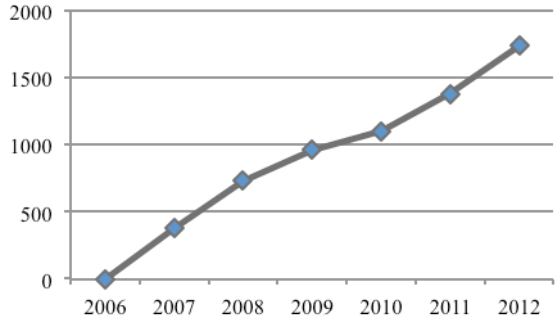


Figure 13a. Number of texts

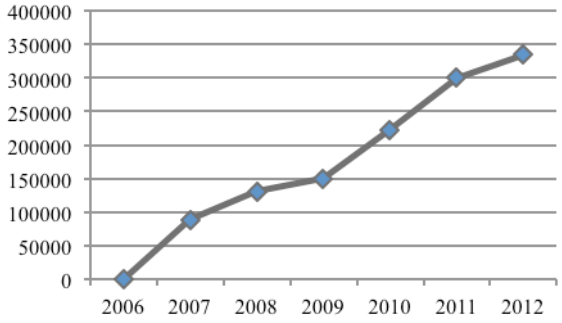


Figure 13b. Number of tokens

Fig. 14 shows the distribution according to genre of the documents written in hieratic script that are encoded and annotated (and the number of documents that await further treatment).[18]

[18] Additionally, more than 400 monumental texts in hieroglyphic script have already been annotated; they represent (a) a selection of 18th dynasty texts whose registers attest evolutionary grammatical features of Late Egyptian; (b) the whole corpus of Ramesside legal decrees; (c) monumental literary texts, like *The Battle of Qadesh*; (d) ideological narratives and rhetorical texts, like the ones of the Medinet Habu inscriptions of Ramses III.

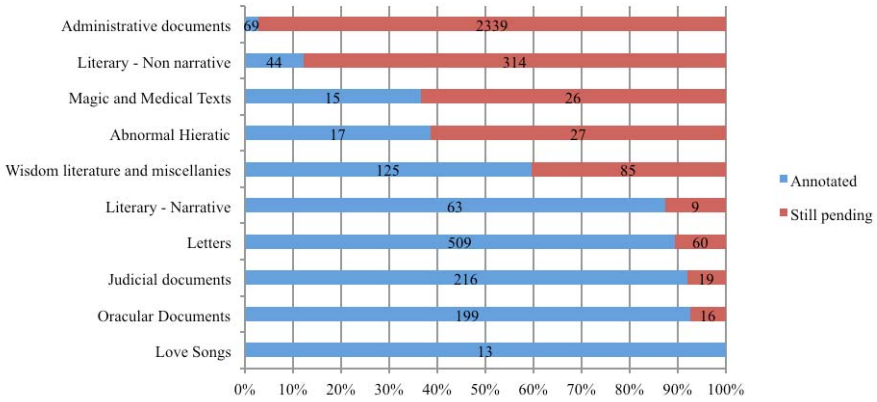


Figure 14. The distribution of hieratic texts according to genres (annotated vs pending)

Given that Ramses is aimed first and foremost at linguistic searches, this figure hardly represents the actual state of the database. Three remarks are warranted here:

(1) Documents deemed more relevant for linguistic analysis have been given high priority. This partially explains the uneven distribution, particularly the small number of administrative documents that have been included in the database up until now.

(2) From the beginning, a deliberate emphasis has been put on the integration of standard editions that contain texts considered to be representative of Late Egyptian: all the texts belonging to the standard collections of texts, such as *LEM*, *LES*, *LRL*, *LRLC*, *RAD*, *TR*, etc. have been completely encoded and annotated.

(3) The length of the documents is highly variable, even within one category. In the category "Wisdom literature and Miscellanies", for examples, among the 85 documents that are still missing, more than 40 are parallel versions on ostraca of the P. Anastasi 1: the longer and/or better preserved documents have been preferred in the first phase of annotation.

Fig. 15a-c show the evolution of the number of lemmata, inflexions, and spellings recorded in the database between 2006 and 2012.

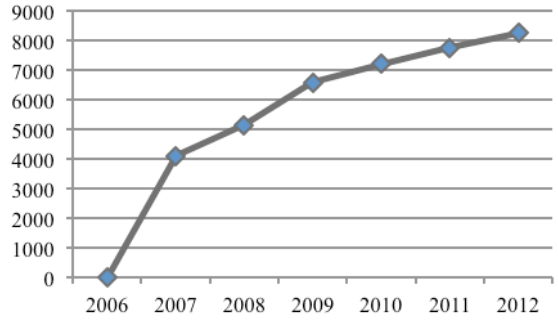


Figure 15a. Number of lemmata

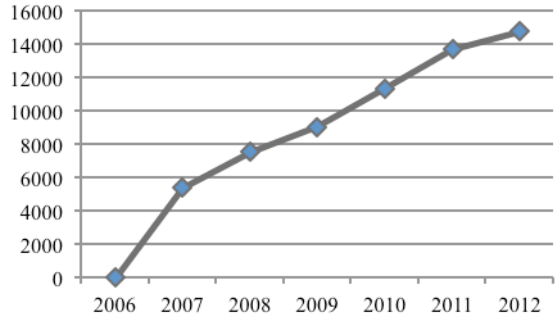


Figure 15b. Number of inflexions

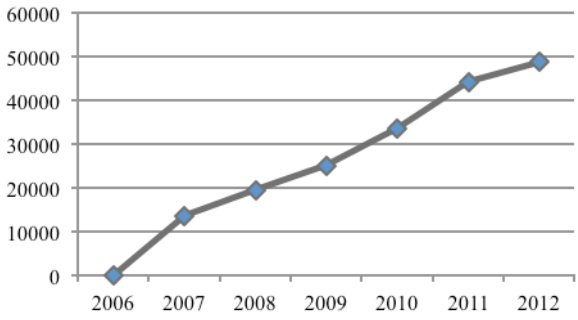


Figure 15c. Number of spellings

As shown in Fig. 15a, the number of lemmata grew quickly during the first year of the project; this results from the fact that the only

dictionary available for Late Egyptian[19] was entirely encoded in the LexiconEditor at the beginning of the project in order to speed up the encoding of the first texts. Otherwise, the progression is regular for the number of inflections and spellings: a bit counter-intuitively, each new text keeps on adding with the same ratio new inflexions and spellings to the database.

2.3 Future perspectives

Before termination of the first phase of the project in October 2013, we will focus on several aspects of Ramses that deserve further attention:

(1) Completion of the encoding and of the annotation of the sub-corpora that we began integrating in Ramses, with a particular focus on the non-narrative literary texts, on the administrative texts and on the texts of the Third Intermediate Period (including the texts written in so-called "abnormal hieratic" or "Kursiv-hieratisch").

(2) New implementations in the TextEditor and SyntaxEditor (ultimately to be merged in a single RamsesEditor). This crucially includes the possibility of defining different levels of access to Ramses (in order to preserve the integrity of the validated data) and a storage of the "history" of successive annotations (when, how and by whom was the annotation carried out? who modified it and when?).

(3) Development of a Web application so as to give the community of Egyptologists and linguists access to the whole range of Ramses data.[20]

Long-term projects include:

(1) The completion of the syntactic annotation of the corpus and the addition of a semantic level of annotation (with word-sense disambiguation).

---

[19] LESKO, L. H., [2]2002-2004: *A Dictionary of Late Egyptian*, 2 vol., 2nd ed., Providence.

[20] We plan to publish the sub-corpora online one after another, immediately after final approval by the team. The end-users will be able to contribute to the enrichment of the corpus thanks to a wiki-like device that will be added in order to allow suggestions regarding the hieroglyphic readings, the addition or emendation of annotations, etc.

(2) The continuation of existing (and development of new) collaborations, e.g. with TXM concerning statistic tools,[21] with the *Thesaurus Linguae Aegyptiae* (see n. 10) in the field of Egyptian lexicography, with the Deir el-Medina Database (see n. 11) regarding the metadata on Late Egyptian texts.

(3) The extension of Ramses' functionalities in order to support earlier and later stages of the Egyptian language, down to Coptic.

3. Changes in methods and practices

The use — the massive use in some cases — of annotated corpora will trigger significant changes in Egyptologists' methods and practices. These changes are, on the whole, indisputably for the better. However, using these new tools without a sharp sense of criticism could potentially lead us in dangerous territories. Here follows a quick review of the main pros and cons.

One of the most obvious advantages of using corpora — even if it is a never ending process — is the exhaustiveness of the data. The textual corpus of Ancient Egyptian (and even a limited subcorpus such as Late Egyptian) is now beyond the reach of a single individual. As one can safely anticipate a regular increase of the data, the benefit of an electronic corpus cannot be overemphasized. Indeed, combined with unlimited numbers of queries on different level of annotation, such corpora should produce falsifiable results in Egyptian linguistics, which is admittedly what is expected from any scientific work.[22]

Electronic corpora, however, could easily give the confidence that they are — even intrinsically, so to speak — objective tools, because they record simple and plain facts. But it would actually be dangerous to assume that databases are neutral from a scientific viewpoint: they are modern ways to organize the rough data. In this respect, Ramses is an annotated corpus, extensively enriched and, as it turns out, choices must be made all the time: in some cases, arbitrary choices that can be explained; in some other cases, choices that are

[21] See http://textometrie.ens-lyon.fr/.

[22] A database like Ramses will make it possible to check hypotheses that unavoidably surface in the course of research projects. This point cannot be overestimated. Scholars are used to the frustrating experience of having failed to take a feature into consideration when reading the corpus. One is then left with two options: neglect it (which can quickly become very problematic from a scientific viewpoint) or start reading the corpus again (which inevitably raises practical problems of time).

the result of the developers' conception of how the grammar of Ancient Egyptian works. In the end, the picture that could emerge from the database is a Late Egyptian grammar *à la liégeoise*, maybe not a bad one in itself, but better to be avoided if one intends to reach a wider audience. To steer clear of such bias, we relied on three strategies aimed at producing a descriptive (i.e. theory-neutral) approach to language structure, with no loss of data because of the resulting method of annotation:

- An analytical approach;
- The possibility of encoding ambiguities;
- The possibility of storing unanalyzable chunks of graphemes.

3.1 An analytical approach to encoding

The choice has been made of coding minimal units rather than larger groups. This is apparent, for instance, in the way lexical composita are handled. In the first place, composita like *mr-mšᶜ* "general" were encoded as one lexical unit. This seemed the most natural way to do it, because it was felt to be very close to every Egyptologist's experience.

This option, however, quickly turned out to be problematic, when less common phrases were to be treated: for instance, is *mr-mšᶜ(-)wr* "general in chief" to be analyzed as a compositum or as two phrases? If *rmṯ-ỉs.t* "crewman" can be safely assumed to be one unit, could this be equally valid for any group with *rmṯ* as its first element? Coptic at first sight seemed to give clear indications,[23] but this turns out to be an illusion. Above all, having large composita would probably have hampered the flexibility of later queries: it would have been impossible, for instance, to look for all the titles containing the qualifier *wr* "great".

Therefore, the decision was finally taken to encode the texts word by word in the TextEditor and to create larger groups with the SyntaxEditor, even in cases where it can safely be assumed that one is dealing with a compositum.

3.2 Encoding ambiguities

Our goal in allowing for the encoding of ambiguities was to lose no piece of information that could be relevant for a query in the corpus.

[23] The Coptic data show that there exist composita built on ⲣⲙ̄- and ⲣⲙ̄ⲙ̄-, which at least suggests different chronological strata.

Accordingly, ambiguities can be encoded at three levels in Ramses (lemma, inflexion, syntactic analysis, and any combination thereof).

Most ambiguities relate to poorly understood contexts, often due to the presence of lacunae. For instance, it is not at all always clear whether a verb is to be analyzed as a perfective or a subjunctive *sḏm=f*. Fig. 16 is an illustration of such a case of morphological ambiguity: in the box of the occurrence, instead of having one analysis, the label <AMBIGUOUS> appears. The two possibilities are recorded in the status line of the word that is displayed at the bottom of the screen (right of Fig. 16). The text can of course been retrieved in any query involving either a perfective or a subjunctive.

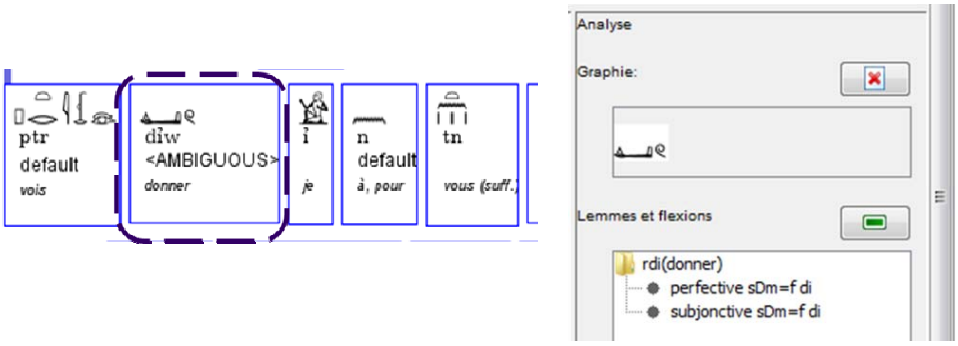


Figure 16. Ambiguities (type 1)

The next example shows another type of ambiguity combining lexical and morphological possibilities (Fig. 17). Due to the fragmentary state of the text, the word *bꜣk* can be understood either as a noun "the work" or as verb "to work". According to the option that will be chosen, the morphological analysis has to be adapted.

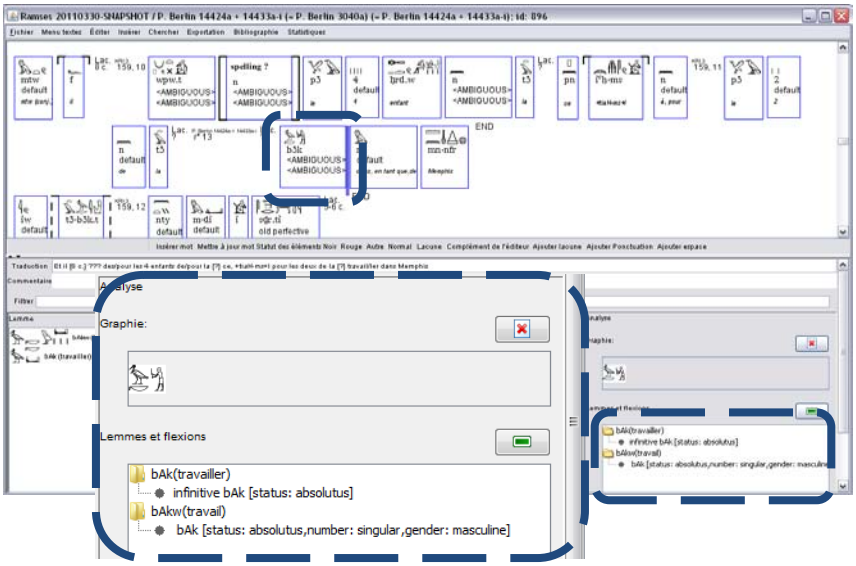


Figure 17. Ambiguities (type 2)

3.3 Encoding unanalyzable sequences of signs

Ramses also makes it possible to encode hieroglyphic signs without linking them to a lemma. This option is of course maximally used in case of lacunae. In doing so, no sign — even if it is completely isolated — is left along the road (Fig. 18).

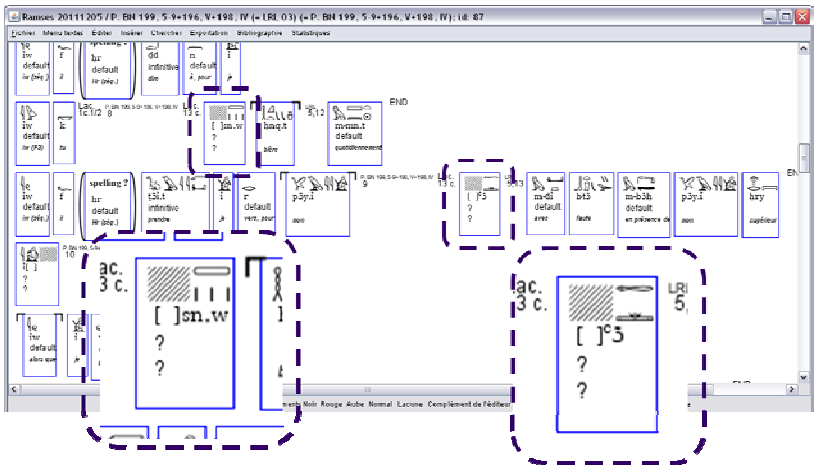


Figure 18. Unanalyzed chunk of graphemes

3.4 Some general reflections

The use of an annotated corpus for data mining seems to offer all possible advantages: it is exhaustive, quick and systematic. But one has to refrain constantly from being naive in the use one makes of an annotated corpus.

First, as remarked above (§3.1), information stored and annotated in the database are never simple facts, directly imported from a supposed objective realm; they always undergo processes of standardization. Second, the extensive — if not sole — use of electronic corpora might entail the risk of discouraging people from developing basic philological skills. There is indisputably some virtue in the old-fashioned habit of reading through whole texts; the exploitation of large corpora solely by means of search engines, even sophisticated ones, usually brings with it a lot of drawbacks, as has become clear for those of us who are accustomed to certain types of typological literature.

Before proceeding to conclusions for this paper, it should be briefly (but plainly) stated what an annotated corpus like Ramses is not, is not yet, and will never be:

(1) Ramses is not a substitute for traditional philological editions. Not only are *facsimile* representations and photographs lacking,[24] but information regarding textual criticism has been kept to a minimum.

(2) Ramses will probably never integrate the vast body of secondary literature that has been written on the texts. In other words, it will never exempt scholars from going back to the secondary literature. As a matter of fact, bibliographical references aim at justifying choices in the annotation (§1.2.2), not at collecting all possible references.

(3) Ramses will never be a substitute — in this case a very bad one to be sure — for a grammar or a dictionary of Late Egyptian. This paper is not the proper place to discuss the all-important issue of lexicographical tools in Egyptology. Some time ago, the *Wörterbuch* team decided that they would not engage a new version of the *Wörterbuch*, but that they would instead provide scholars with an electronic thesaurus, the *Thesaurus Linguae Aegyptiae* (see

[24] From a technical point of view, this issue can be very easily addressed, but problems regarding the copyrights and credits for the pictures are still to be dealt with.

n. 10). As a consequence, it is now up to everyone to write his/her own lexicographical notes based on the data of the *TLA*, which is a complete change of paradigm. On the contrary, in our eyes, Egyptologists need a proper and modern dictionary. A new dictionary of Late Egyptian is thus one of the major achievements that could be produced with the help of Ramses, but the output will clearly be outside the scope of the Ramses project.[25]

4. Conclusions: New anvenues for research

Notwithstanding the foregoing observations, annotated corpora like the *TLA* or *Ramses* will bring significant positive changes in the study of the Ancient Egyptian language(s) and texts.

The SearchEngine under development in the framework of Ramses will indeed not only make queries far easier to execute than ever before, but — above all — it will allow queries that could not have been previously achieved on account of the high degree of complexity and/or the size of the corpus to be examined. By way of conclusion, we will point to some research domains that were, on the one hand, already accessible with traditional tools but that can now be approached faster, more systematically, and more exhaustively and, on the other hand, new avenues for research that were simply impossible to pave without such richly annotated corpora.

In the sphere of traditional philology, a corpus like Ramses could help considerably in taking up the challenge of the identification and grouping of hundreds of pieces of literary texts on ostraca that are scattered in collections and museums all over the world. If one is faced, for example, with the sequence of graphemes  , the identification of the text (even if well-known) is a long and uncertain endeavor. A simple query in Ramses — that is built according to the most probable segmentation for this sequence of graphemes (see Fig. 19) — gives two results, both from copies of the *Teaching of Amennakhte*.

[25] This issue has been discussed by Winand in a conference held in Leipzig in November 2012.

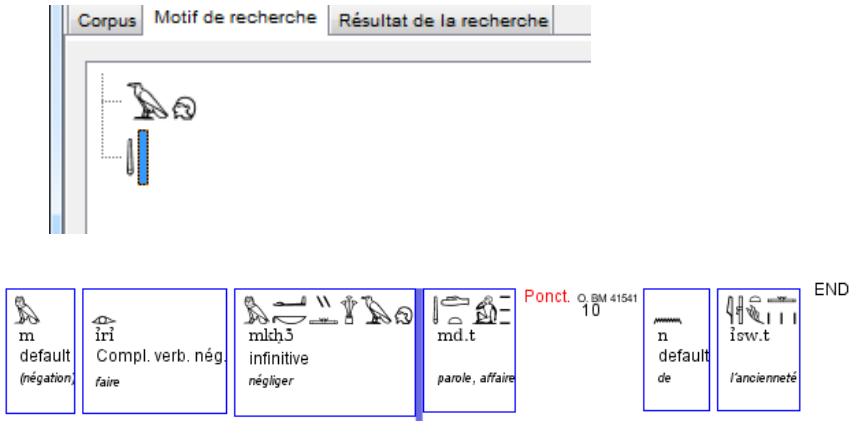


Figure 19. The identification of literary fragments

In the domain of graphemics, researches on the system of classifiers can be broached much more easily. For instance, listing all the lexemes that can have the 𓇋 classifier was a long, fastidious and possibly non exhaustive task without an annotated corpus, while Ramses produces a list of 53 related lexemes in the corpus instantaneously. More problematic would be studies that involve the combination of the graphemic level with other level of analysis. One can think, for example, of the combination of the divine classifier 𓇓 with pronominal elements. A query like the one of Fig. 20 gives directly 1358 matches that can be sorted according to any kind of criterion.

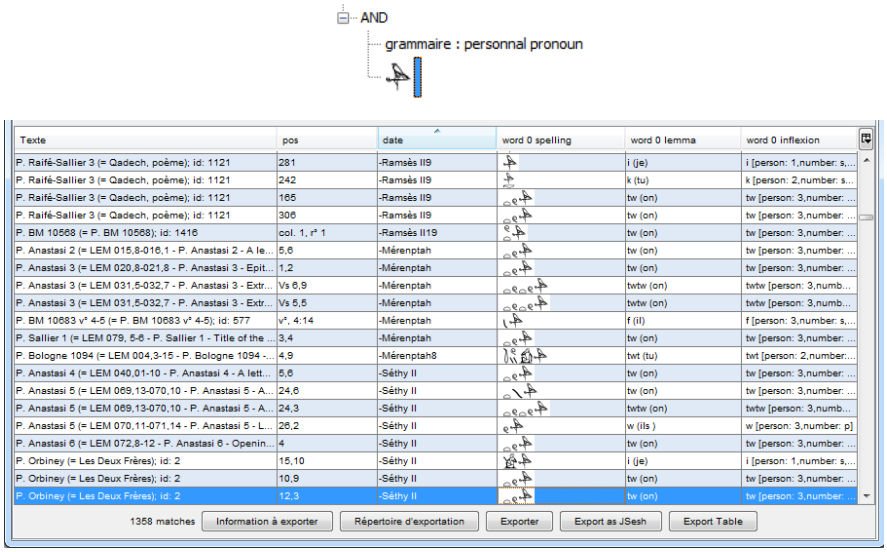


Figure 20. Research on classifiers

If the benefits of an annotated corpus in the field of morphology and syntax are obvious, it should be stressed that the use of semantic information that are stored in the LexiconEditor in combination with morphosyntactic features opens new opportunities for checking hypotheses, e.g. about (the evolution of) the selectional restrictions of constructions. The query of Fig. 21 in the TextEditor, for instance, allows one to find all the occurrences of Future III with inanimate subjects.

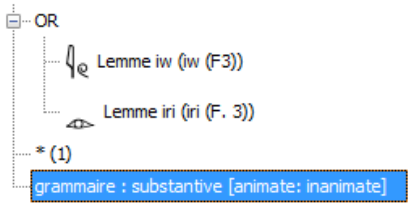


Figure 21. Types of subject with the Future III: Inanimate subjects

Finally, Natural Language Processing — an entirely new field for Egyptology — will be explored in close collaboration with computer scientists. The first applications that come to mind are: the develop-

ment of taggers and parsers, automatically generated indexes and concordances, the application of methods for automatic text categorization (e.g. with decision trees) and information retrieval, as well as advanced statistical tools.