COLLECTION
ÆGYPTIACA LEODIENSIA

La collection *Ægyptiaca Leodiensia* — dirigée par Jean Winand, Dimitri Laboury et Stéphane Polis — a pour vocation de publier des travaux d'égyptologie dans les domaines les plus divers. Elle accueille en son sein des monographies ainsi que des volumes collectifs thématiques.

This volume represents the outcome of the meeting of the Computer Working Group of the International Association of Egyptologists (Informatique & Égyptologie) held in Liège in 2010 (6-8 July) under the auspices of the Ramses Project. The papers are based on presentations given during this meeting and have been selected in order to cover three main thematic areas of research at the intersection of Egyptology and Information Technology: (1) the construction, management and use of Ancient Egyptian annotated corpora; (2) the problems linked to hieroglyphic encoding; (3) the development of databases in the fields of art history, philology and prosopography. The contributions offer an up-to-date state of the art, discuss the most promising avenues for future research, developments and implementation, and suggest solutions to longstanding issues in the field.
Two general trends characterize the projects laid out here: the desire for online accessibility made available to the widest possible audience; and the search for standardization and interoperability. The efforts in these directions are admittedly of paramount importance for the future of Egyptological research in general. Indeed, for the present and increasingly for the future, one cannot overemphasize the (empirical and methodological) impact of a generalized access to structured data of the highest possible quality that can be browsed and exchanged without loss of information.

Stéphane POLIS is Research Associate at the National Fund for Scientific Research (Belgium). His fields of research are Ancient Egyptian linguistics and Late Egyptian philology and grammar. His work focuses on language variation and language change in Ancient Egyptian, with a special interest for the functional domain of modality. He supervises the development of the Ramses Project at the University of Liège with Jean Winand.

Jean WINAND is professor ordinarius at the University of Liège, and currently Dean of the Faculty of Philosophy and Letters. He specializes in texts and languages of ancient Egypt. His major publications include *Études de néo-égyptien. La morphologie verbale* (1992); *Grammaire raisonnée de l'Égyptien classique* (1999, with Michel Malaise); *Temps et Aspect en égyptien. Une approche sémantique* (2006). He launched the Ramses Project in 2006, which he supervises with Stéphane Polis.

PRESSES UNIVERSITAIRES DE LIÈGE

ISBN : 978-2-87562-016-3

ÆGYPTIACA LEODIENSIA

9

Texts, Languages & Information Technology in Egyptology
Stéphane POLIS – Jean WINAND

COLLECTION
ÆGYPTIACA LEODIENSIA

Texts, Languages & Information Technology in Egyptology

Stéphane POLIS — Jean WINAND
With the collaboration of Todd GILLEN

Presses Universitaires de Liège

Texts, Languages & Information

Technology in Egyptology

Dépôt légal D/2012/12.839/17
ISBN 978-2-87562-016-3

Place du 20-Août, 7
B-4000 Liège (Belgique)
http://www.presses.ulg.ac.be

Imprimé en Belgique

Collection *Ægyptiaca Leodiensia* 9

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

Stéphane POLIS & Jean WINAND (eds.)

With the collaboration of Todd GILLEN

Presses Universitaires de Liège
2013

Table of Contents

Stéphane POLIS, Texts, Languages & Information Technology in Egyptology. Introduction ..... p. 7-10

1. *Annotated corpora of Ancient Egyptian texts*

Peter DILS & Frank FEDER, The *Thesaurus Linguae Aegyptiae.* Review and Perspectives..... p. 11-23

Stéphane POLIS, Anne-Claude HONNAY & Jean WINAND, Building an Annotated Corpus of Late Egyptian. The Ramses Project: Review and Perspectives..... p. 25-44

Stéphane POLIS & Serge ROSMORDUC, Building a Construction-Based Treebank of Late Egyptian. The Syntactic Layer in Ramses..... p. 45-59

Stéphanie GOHY, Benjamin MARTIN LEON & Stéphane POLIS, Automated Text Categorization in a Dead Language. The Detection of Genres in Late Egyptian ..... p. 61-74

Mark-Jan NEDERHOF, Flexible Use of Text Annotations and Distance Learning ..... p. 75-88

2. *Hieroglyphic encoding*

Roberto GOZZOLI, Hieroglyphic Text Processors, Manuel de Codage, Unicode, and Lexicography..... p. 89-101

Mark-Jan NEDERHOF, The Manuel de Codage Encoding of Hieroglyphs Impedes Development of Corpora..... p. 103-110

Vincent EUVERTE & Christian ROY, Hieroglyphic Text Corpus. Towards Standardization..... p. 111-120

3. *Databases for art history, texts and prosopography*

Christian MADER, Bernhard HASLHOFER & Niko POPITSCH, The MEKETREpository. A Collaborative Web Database for Middle Kingdom Scene Descriptions..... p. 121-128

Nathalie PRÉVÔT, The Digital Puzzle of the *talatat* from Karnak. A Tool for the Three-Dimensional Reconstruction of Theban Buildings from the Reign of Amenhotep IV..... p. 129-138

Carlos GRACIA ZAMACONA, A Database for the Coffin Texts ..... p. 139-155

Azza EZZAT, The Digital Library of Inscriptions and Calligraphies..... p. 157-161

Yannis GOURDON, The *AGÉA* Database Project.
Anthroponymes et Généalogies de l'Égypte Ancienne .....p. 163-168

Eugene CRUZ-URIBE, Computers and Journal Publishing. A Position Paper .....p. 169-174

Abstracts.....p. 175-178

Building an Annotated Corpus of Late Egyptian*

The Ramses Project: Review and Perspectives

Stéphane POLIS[§], Anne-Claude HONNAY & Jean WINAND

F.R.S.-FNRS[§] – Université de Liège

1. INTRODUCTION

The Ramses project aims at building a richly annotated historical corpus[1] of all Late Egyptian texts and, more broadly, of all the written material whose linguistic registers attest Late Egyptian linguistic features from the 18th dynasty down to the Third Intermediate Period (ca. 1350-700 BCE). The database will ultimately include, for each text, all 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, i.e. sub-corpus by sub-corpus, provide online access to the Ramses annotated corpus.

Since the beginning of the project in July 2006, two reports have been made:[2] in the first one,[3] after an overview of existing lexical databases and annotated corpora in Egyptology, we focused on the motivations for launching such a project and on the available human resources and presented a beta-version of the IT developments that needed to be fully implemented in order to facilitate the encoding of hieroglyphic and hieratic texts; in the second one,[4] we described more precisely the process of text encoding in Ramses (TextEditor and LexiconEditor) and addressed the kind of functionalities implemented in the Search Engine.

The Ramses database has developed and improved in many respects since these reports. In the present paper, we first review different aspects of the project: presentation of the current scientific team (§2.1) and of the progress made in the encoding of the corpus (§2.2); general description of the software that is currently fully operational (§2.3). In a second section, we introduce two new functionalities that have recently been incorporated in the Ramses software: RamsesBib, a tool for handling bibliographical information at every level (§3.1) and RamsesExport, a tool for exporting data (§3.2). Finally, we set out the latest version of Search-Engine (§4) that has been subject to considerable deve-

* We are grateful to Todd Gillen for several suggestions regarding the content of this paper as well as for editing the English with his usual competence. The data of this paper have been collected on 2011/12/31.

1. "[…] a 'historical corpus' is one which is intentionally created to represent and investigate past stages of the language and/or to study language change." (Claridge 2008: 242).
2. Additionally, a first note of intention is to be found in Polis 2006.
3. A lecture delivered at the 10th Congress of the International Association of Egyptologists in Rhodos (May 2008), see Winand, Polis & Rosmorduc (in print).
4. Proceedings of the *Informatique & Égyptologie* meeting in Vienna (July 2008), see Rosmorduc, Polis & Winand 2009.

lopment in the last few years: its nearly unlimited potential in terms of types of queries promises to open up new avenues of research for Egyptian linguistics.

2. CURRENT STATE OF THE PROJECT

Building a large (over 1 million words) and richly annotated corpus — graphemics, morphology, syntax, semantics as well as meta-data (corpus mark-up) are recorded — of a language such as Late Egyptian calls for considerable human resources and interdisciplinary collaboration (§2.1). Indeed, no previous electronic data were available; moreover, unlike in most of the languages dealt with in corpus linguistics, intensive philological preparation is needed for every single document prior to encoding. The digitization of the whole corpus is performed manually: encoding of hieroglyphs, lemmatization and part of speech tagging, syntactic parsing as well as metadata collection (§2.2 for the progress in the encoding). Therefore, the process of encoding and annotating the corpus had to be facilitated by a software solution (§2.3) that would guarantee speed of application and ensure the coherence and consistency of the analyses.

2.1. *Scientific team and collaborations*

The project, started in 2006, is carried out under the academic supervision of Jean Winand (University of Liège). The scientific direction is jointly assumed by Jean Winand and Stéphane Polis (F.R.S.-FNRS – University of Liège) and IT developments are made under the supervision of Serge Rosmorduc (Paris-VIII – Conservatoire National des Arts et Métiers).

The project is principally funded by the University of Liège (a five year program called 'Action de recherche concertée [ARC]') and by the F.R.S.-FNRS (a four year stipendium called 'Fonds de la recherche fondamentale collective [FRFC]'). Some younger scholars who benefit from a doctoral fellowship also work on Ramses. Since the beginning of the project, the team (ten persons nowadays) has undergone some transformations. The following list attempts to keep track of this evolution:

- J. Winand, St. Polis and S. Rosmorduc have been a part of the project from the beginning;
- L. Neven worked for the project from 2006-2010 as a doctoral fellow with the ULg;
- A.-Cl. Honnay, also a team member from the outset, now works with a grant from the FRFC;
- St. Gohy joined the team in 2008 as a junior fellow with the F.R.S.-FNRS;
- A. Stella benefits from a doctoral grant funded by the ARC (2008-2012); this was also the case of J. Raimondo, who left the project in 2011 and has been replaced by Guillaume Lescuyer; B. Martin Leon also funded by the ARC has a masters degree in computer engineering, he assists S. Rosmorduc in the IT developments and he currently undertakes a PhD dissertation on semi-automatic tagging and parsing of Late Egyptian texts;
- N. Sojic joined the team in October 2011 as a doctoral fellow with the ULg;
- A.-L. Comhaire (MA in Egyptology) works to encode texts as a volunteer.

Since 2008, we are able to fund post-doc students to help us in developing specific parts of the project:

- In 2008-2009, D. Lefèvre (ÉPHÉt.-Paris, now University of Geneva) assisted in systematizing some parts of the lexicon (esp. the titles and composita) and encoded the el-Hibeh letters,
- In 2009-2010, E. Grossman (University of Jerusalem) intensively worked on the principles of the SyntaxEditor with St. Polis; he is now a Martin Buber fellow in Jerusalem,
- Since 2010, T. Gillen (Macquarie University) assists in double-checking the texts already encoded; he also takes charge of the Medinet Habu inscriptions, and more generally of the epigraphic material of the Ramesside period.

Additionally, the Ramses team has developed over the years collaborations with other projects and scholars, especially in the fields of syntactic analysis and text corpus statistics:

- The LASLA[5] ("Laboratoire d'analyse statistique des langues anciennes", ULg) that is working on the implementation of a syntactic parser for Latin texts.
- Nicolas Mazziotta (ULg) who developed the open source *Notabene* software,[6] a tool designed for multiple linguistic annotations of text corpora.
- *Unitex*,[7] a corpus processing system based on automata-oriented technology developed chiefly by Sébastien Paumier (University of Paris-Est Marne-la-Vallée).
- The research project Textométrie[8] that developed TXM, a platform which combines powerful techniques for the analysis of large bodies of texts into a modular and open-source framework. In this context, particularly worth mentioning is the fact that Serge Heiden and Alexei Lavrentiev (ENS-Lyon) designed a TEI-compatible (http://www.tei-c.org/index.xml) XML mark-up pivot format allowing Ramses data to be imported into the TXM platform; this will ultimately give Ramses users access to the powerful statistical capabilities of TXM that are based on R[9] (a language an environment for statistical computing and graphics).

2.2. *The encoding: texts, words, lemmata, inflections and spellings*

Whatever may be the quality of the tools developed for facilitating the encoding (see §2.3), Ramses is a completely manually annotated corpus, which means that the process of integrating a text in the database is somewhat tedious and undoubtedly time consuming. In order to overcome the problem or, at least, to limit the inconvenience, we devised two strategies for the enterprise:

(1) The implementation of user-friendly software (see §2.3) facilitating the data capture (including the hieroglyphic script).
(2) The splitting of the corpus into sub-corpora according to genres and period. Indeed, the written registers of Late Egyptian are highly diverse in terms of lexicon, phraseology, distribution of inflectional patterns, etc. The choice was thus made to divide the corpus between annotators; this is intended (a) to speed up the process of annotating and (b) to increase the coherence of the encoding (at least with recurrent patterns).

Currently, more than 1350 texts (see Fig. 2e) have been included in the database and received multifaceted annotations. Fig. 1 shows the distribution of the documents (written in hieratic script[10]) that are encoded and annotated (and the number of documents that await further treatment):

5. http://www.cipl.ulg.ac.be/Lasla/index.html.
6. See Mazziotta 2010a & 2010b.
7. http://igm.univ-mlv.fr/~unitex/.
8. http://textometrie.ens-lyon.fr/?lang=en.
9. http://www.r-project.org/.
10. Additionally, more than 400 monumental texts (hieroglyphic script) have already been annotated; they represent (a) a selection of 18th dynasty texts whose registers attest evolutionary grammatical features of Late Egyptian (this includes, *inter alia*, various texts from the Amarna period), (b) the whole corpus of Ramesside legal decrees (see David 2006), (c) monumental literary texts, like *The Battle Qadesh* of Ramses II, (d) ideological narratives and rhetorical texts, like the Medinet Habou inscriptions of Ramses III.

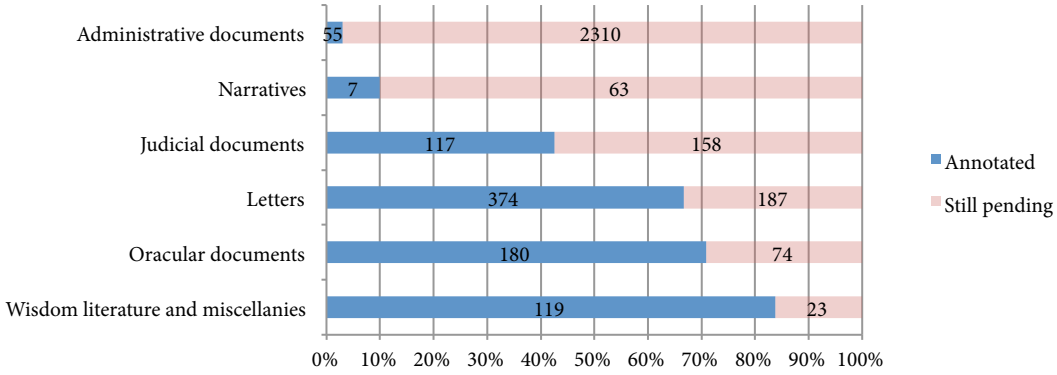

Figure 1. Hieratic documents annotated in Ramses

Given that Ramses is aimed first and foremost at linguistic searches, Fig. 1 hardly represents the actual state of the database, and several remarks are warranted in this respect:

(1) Documents deemed more relevant for linguistic analysis have been favoured. 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, an emphasis has been put on the integration of standard editions that contain texts deemed to be representative of Late Egyptian. Therefore, all the texts belonging to the *LEM*,[11] *LES*, *LRL*, *LRLC*, *RAD*, *TR* have been completely encoded and annotated.

(3) The length of the documents is highly variable, even within one category: among the narratives, for example, the number of annotated documents (*LES*) constitutes less than 10% of the extant documents preserving narrative literary texts; however, these texts represent more than 70% of the corpus in terms of tokens or "words". The longer and better preserved documents have been preferred in the first phase of annotation.

Figs. 2a-e show the evolution of the number of, respectively, lemmata, inflections, spellings, words and texts recorded in the database between 2006 and 2011.

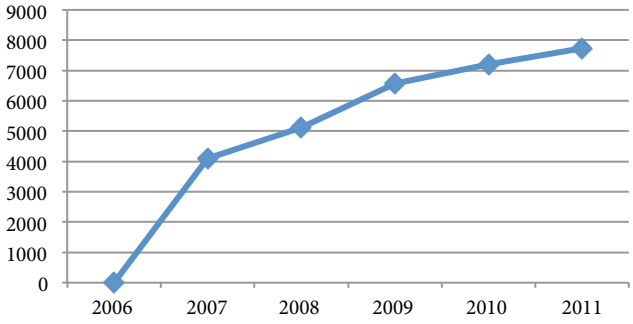

Figure 2a. Lemmata

11. For these abbreviations, see the bibliography.

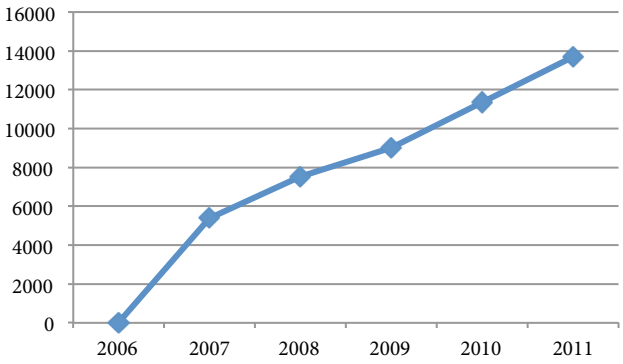

Figure 2b. Inflections

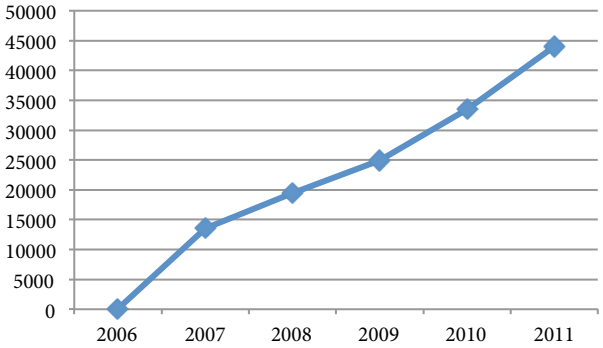

Figure 2c. Spellings

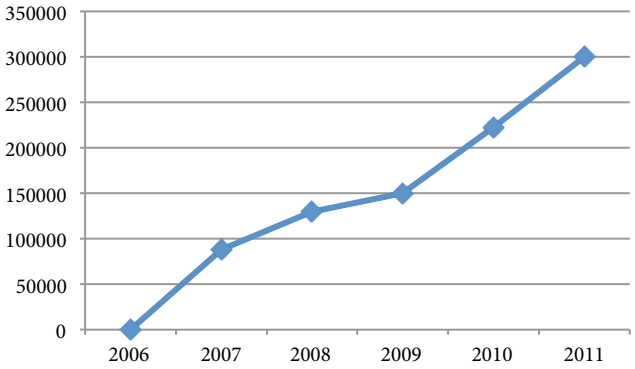

Figure 2d. Words

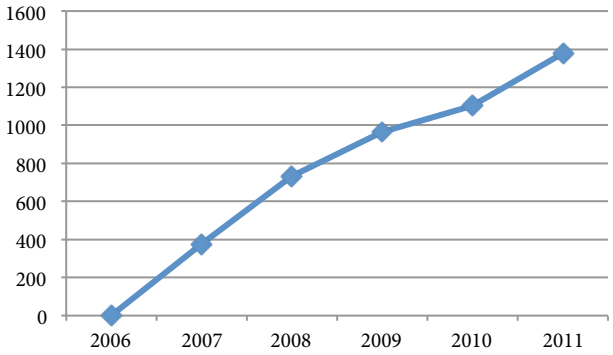

Figure 2e. Texts

As shown by Fig. 2a, the number of lemmata grew quite quickly during the first year of the project; this results directly from the fact that, at the beginning of the project, the only dictionary available for Late Egyptian[12] was entirely encoded in the LexiconEditor so as to be on firm ground for the encoding of the first texts. Otherwise, as shown by Figs. 2b-e, the progression is quite regular (and parallel) for the number of inflections, spellings, words (nearly 300 000 words) and texts; the last two years even testify a slight increase of the number of new words annually annotated in the database (Fig. 2d), which has resulted from capitalization on the strong base of a well-stocked LexiconEditor.

2.3. *The Ramses software*

As a manually annotated corpus constituting, from a technical point of view, a relational database in SQL where the texts are represented and stored in XML, Ramses had to meet two types of basic requirements:

(1) 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 (if not accuracy) of annotation.
(2) From the user's point of view, the annotation schemes should allow for an extreme sensitivity of analysis, but also avoid adherence to any strict theoretical linguistic framework, so as to allow for a wide range of end-users (see Leech 2003).

In order to meet the annotator's requirements in terms of speed and consistency, two interrelated JAVA modules have been designed for handling the graphemic and morphological levels: a TextEditor and a LexiconEditor. The principle at work is the following: each occurrence of a word in a text (TextEditor) is the actuation of a detailed entry in the lexicon (LexiconEditor). In other word, in the process of encoding a text in the TextEditor, the encoder simply has to select the appropriate lemma, inflection and spelling in constrained lists (bottom part of the screen) that summarize the data already encoded in the LexiconEditor (see Fig. 3).

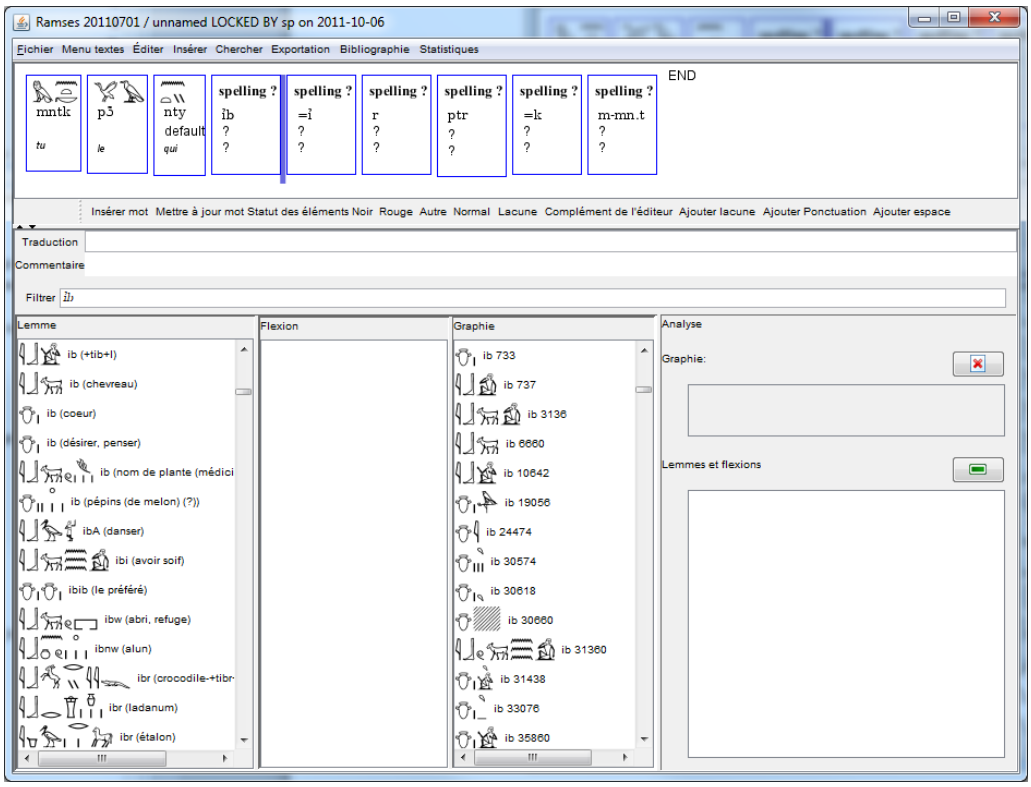

Figure 3. TextEditor: Enhancing the speed of annotation

12. Lesko [2]2002-2004; each entry has been checked against the *Wörterbuch der ägyptischen Sprache*.

If any lemma, inflection or spelling is missing, these lists can be supplemented by adding new information in the LexiconEditor (see Fig. 4). As one can imagine, the encoding of texts was quite slow in the beginning; but with the growth of the corpus and the expansion of the data in the LexiconEditor, the annotator's work becomes correlatively faster.

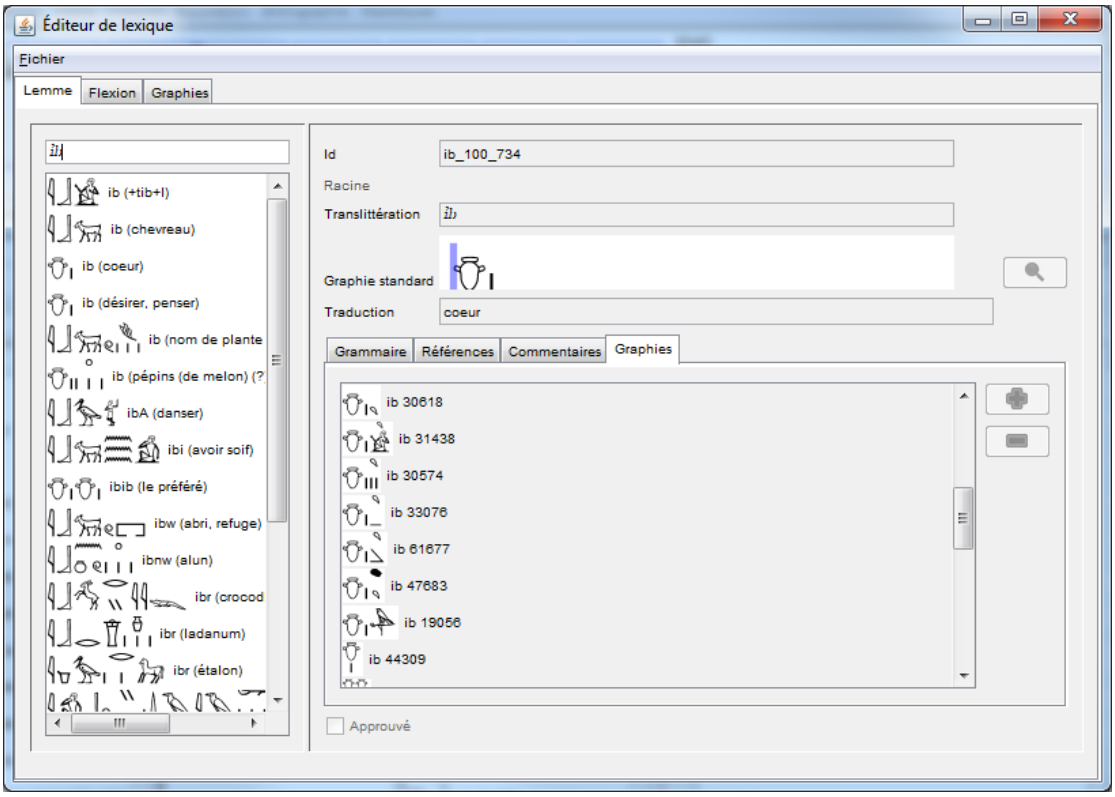

Figure 4. LexiconEditor: Some of the spellings attested for *ib* "heart"

In this respect, the next step will consist in the implementation of a context-sensitive semi-automatic tagger that suggests to the annotator the lemma, inflection and spellings that are the most likely for a word while taking into account mark-up data such as the genre, date and support of any new text.

On the other hand, so as to cope with the user's need for fine-grained data and detailed linguistic analyses, the number of levels of annotations in Ramses is maximal. At the same time, the tags and labels are intended to be linguistically consensual, i.e. as purely descriptive as possible in order to keep the database free from any specific formalism.[13] The annotations in Ramses may be subsumed under three main headings: (1) corpus mark-up, (2) ecdotic descriptors and (3) linguistic annotations:

(1) Ramses includes corpus mark-up, i.e. meta-data about the texts (genre, linguistic register, etc.) and documents (date, nature of the writing support, writing system, place of origin, etc.).[14] This allows for a wide range of questions to be explored, most importantly sociolinguistic (dialects, registers, etc.) and diachronic variation.

(2) As a text language, Late Egyptian has come down to us only through (usually fragmentary) documents — ostraca, papyri, tablets, stelae or inscribed walls. Additionally, at the risk of stating the obvious, no (native speaker) informant can be asked to clear up an obscure

13. Of course, even apparently basic matters, such as defining a set of word classes (POS), are possibly subject to disagreement (see infra in §2.3 for a possible answer to the critics of using categories that have been developed in pre-corpus days).

14. In the near future, we plan to include additional metadata about the name of scribes and copyists where identification with historical figures have been proven or suggested; this should lead to entirely new types of variationist approaches to the Ancient Egyptian language.

sentence or to account for unexpected constructions. Accordingly, the philological dimension had to be fully taken into account within Ramses. This resulted in three decisions: (a) textual criticism is entirely integrated (see Fig. 5) with specific tags referring, on the one hand, to the actual state of preservation of the documents (lacuna, erasure, etc.) and to scribal peculiarities (*supra/infra lineam* addition, etc.) and, on the other hand, to the philological editing of texts (editor's emendation, addition, etc.); (b) bibliographical information can be linked to any type of annotation in order to justify the choices and interpretations based on the extant literature in the field; (c) annotators are never forced to opt for an annotation (see Fig. 6) if the context and/or actual state of preservation of the document does not allow for choosing one reasonably: spellings may be added without them being linked to any given lemma or inflection, a word may be lemmatized with no inflectional analysis, etc.

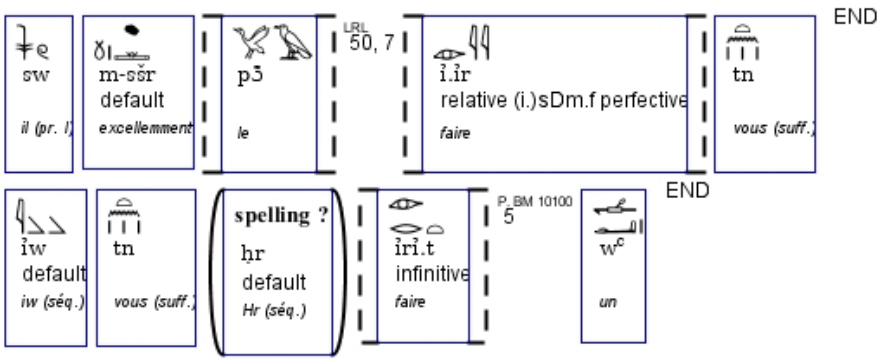

Figure 5. Visualization of tags relating to textual criticism in the TextEditor

(3) Linguistic annotations are independent from the graphemic level and added using XML mark-up language, so that no integrity of the data is lost in the process of enriching the corpus.[15] Moreover, in order not to freeze the information by imposing one particular linguistic analysis on one annotation, the coding of ambiguity is fully supported by Ramses (see Fig. 6): each sequence of hieroglyphs can be assigned to several lemmata and/or inflections if various analyses suggest themselves to the annotator. As for the content of these annotations, the linguistic tagging is not guided by specific types of linguistic exploitations, but it should ideally be able to produce results for any kind of research. Therefore, data regarding various levels of analysis concerning the lemmata (root, part-of-speech, morphological class, valency, semantic class, etc.) and the inflections (all the morphological patterns) can be specified.[16]

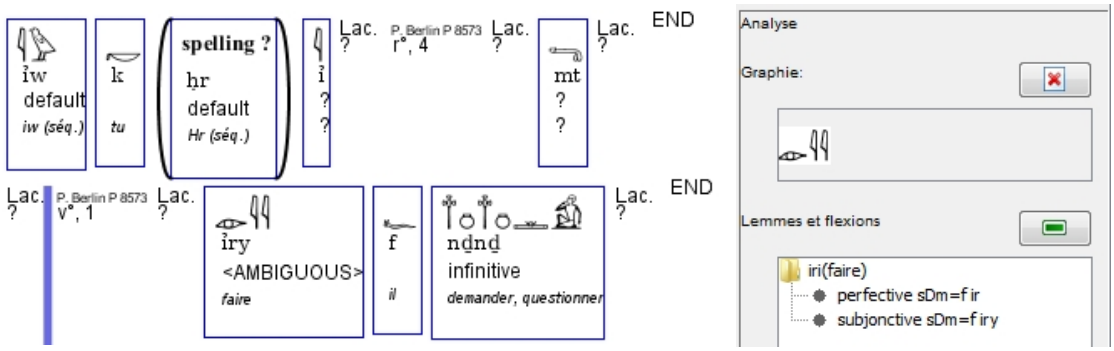

Figure 6. Underspecified annotations and coding of ambiguity in the TextEditor

15. Additionally, whenever needed, the capabilities of the Search-Engine (see §4) make it possible to ignore any level of annotation. Consequently, annotations never clutter up the data.

16. The annotation scheme is based on guidelines which are described in a "Manuel d'encodage" (Honnay & Polis 2011) and will ultimately be available online to end-users.

The syntactic annotation of the corpus, now made using the SyntaxEditor, is still in a test phase (see Polis & Rosmorduc in the current volume).

The functionalities of the SyntaxEditor have been developed in order to allow not only for phrasal chunking 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, e.g. with the co-indexation of pronouns and noun phrases) and information structure as well as speech acts.

The annotation scheme is a priori neither framed in a constituent structure model nor in a dependency model, for we envision these representations as two different, but nevertheless possible, outputs of a single 'construction-based' syntactic annotation. The syntactic scheme has been (and continues to be) developed in order to account for the diversity of linguistic facts found in the Late Egyptian corpus; it takes seriously the assumption of Construction Grammar that *constructions* are the basic units of 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 (e.g. part-of-speech) categorization (as annotated in the TextEditor). This means that syntactic annotation will undoubtedly have a feed-back effect on the previous analyses, thereby avoiding the methodologically untenable position (see e.g. Hunston 2002: 93) of a priori defining a category such as part-of-speech.

From an IT point of view, the TextEditor and the SyntaxEditor will eventually merge into a single JAVA module with visualization facilities that will enable the annotators to select the level of linguistic analysis to which they wish to have access.

3. TWO NEW FUNCTIONALITIES: RAMSESBIB AND RAMSESEXPORT

Two new functionalities have recently (2010-2011) been implemented in Ramses: RamsesBib, a tool for handling bibliographical information (§3.1) and RamsesExport, a tool for exporting data (§3.2).

3.1. *RamsesBib*

We have integrated into Ramses the rich and abundant modern literature on Late Egyptian texts and lexemes in a principled way, designed not only to meet the philological requirements of Ancient Egyptian linguistics, but also to make explicit the analytical choices made during annotation of the data: end-users should easily understand the reasons for preferring one analysis to another. Ramses aims not only at building an annotated corpus, but also at eventually collecting all the references that may be relevant to the study of Late Egyptian texts.

To these ends, a specific module, called RamsesBib, has been implemented by Benjamin Martin Leon (ULg; Ramses) and Laurent Simon (ULg; Centre Informatique de Philosophie et Lettres [CIPL]). It can be accessed directly via the main menu of the TextEditor (see Fig. 7).

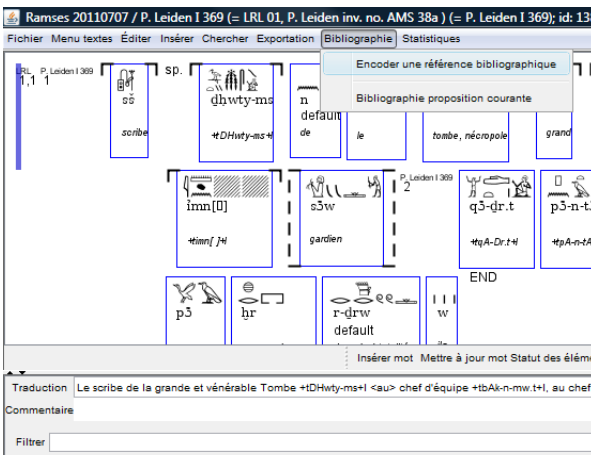

Figure 7. Adding new bibliographical references

Within the main tab of RamsesBib (see Fig. 8), every kind of bibliographical reference may be stored in the database.[17] In order to ensure maximal consistency in the encoding of new references, only the field "title" and those concerning editorial data can be filled out freely. The other fields (author, collective work, journal, series) are drop-down lists which can be enriched via other tabs in the RamsesBib module.

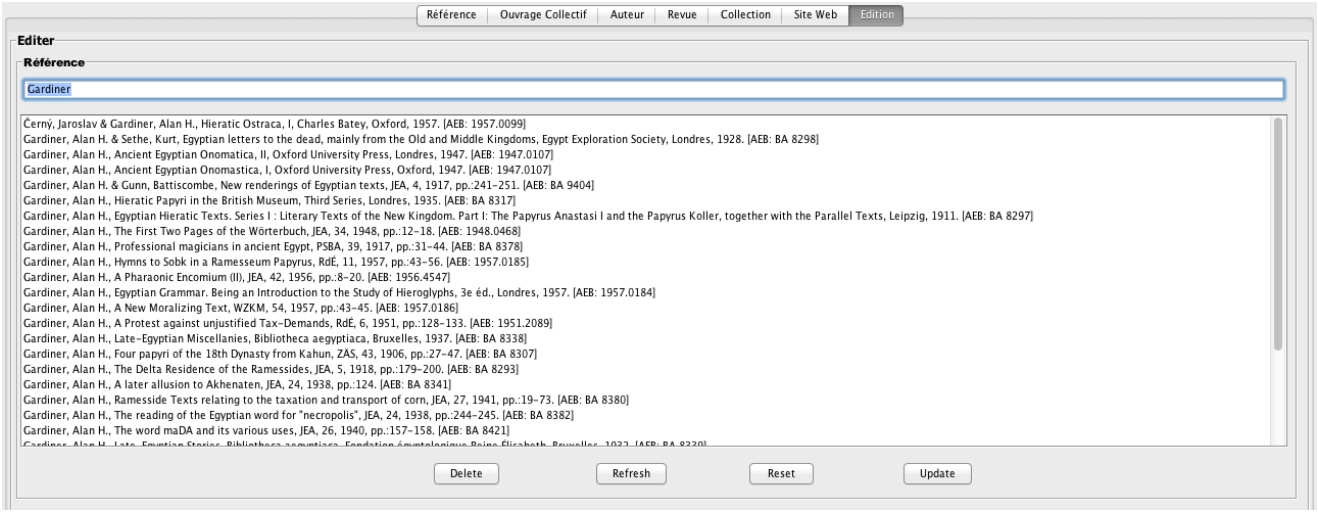

Figure 8. List of references containing the string "Gardiner" encoded in RamsesBib

The digital era sees an increasing number of resources available on the Web. Therefore, a specific tab is dedicated to the encoding of institutionally supported web sites that publish online textual and/or lexicographical resources[18] as well as meta-data concerning the Late Egyptian corpus.[19] Once encoded, all these data are directly accessible online to Ramses end-users.

A last tab of RamsesBib gives access to the full list of bibliographical references recorded in the database and allows editing and emendating.

After it has been encoded in RamsesBib, any reference can be instantiated in different parts of the TextEditor and LexiconEditor (see §2.3). Within the LexiconEditor, bibliographical information can be added at three levels:

(1) the lemma (see Fig. 11),
(2) the inflection,
(3) the spelling.

Within the TextEditor, references can be linked to:

(1) the description of a text (see Fig. 9),
(2) any proposition in a text (this is meant to include in the database the references to passages quoted and discussed in grammars and individual studies).

17. Especially noteworthy is the addition, for each bibliographical entry, of the Online Egyptological Bibliography code (see http://oeb.griffith.ox.ac.uk/). This is eventually meant to allow Ramses end-users to access online the references and abstracts in the OEB directly (see http://oeb. griffith.ox.ac.uk/).

18. E.g. Deir el Medine Online (see http://dem-online.gwi.uni-muenchen.de/) or the *Thesaurus Linguae Aegyptiae* (see http://aaew.bbaw.de/tla/, see the review in the current volume).

19. For example the Deir el-Medina Database (http://www.leidenuniv.nl/nino/dmd/dmd.html).

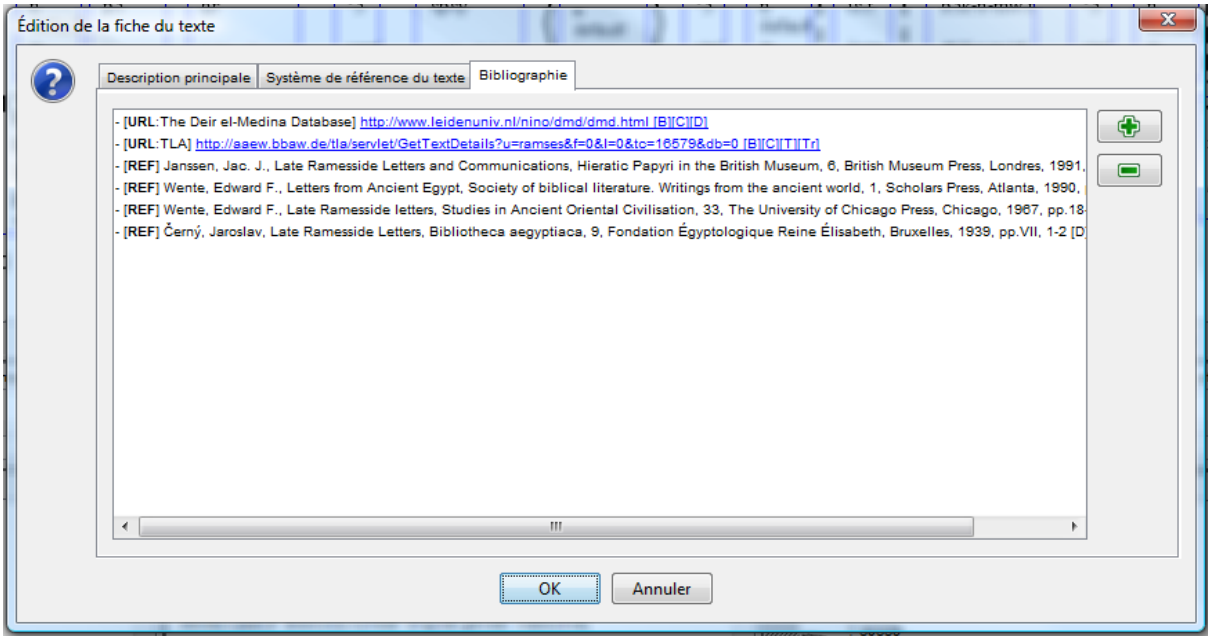

Figure 9. (Selective) bibliography of a text

For each actuation of a reference in the database, the encoder can not only specify the pages and figures concerned (and possibly add some comment), but also tag the content of the quoted bibliographical entry. This practice was directly inspired by the *TLA*,[20] where each kind of content occurring in a reference is identified by an acronym: bibliography [B], commentary [C], description [D], facsimile [F], photograph [P], hieroglyphic transcription [H], transliteration [T] or translation [Tr]. This functionality enables end-users to generate automatically lists of references regarding a specific aspect of a text (e.g. 'list of all the hieroglyphic editions', etc.), of a lemma, or even of a sentence.

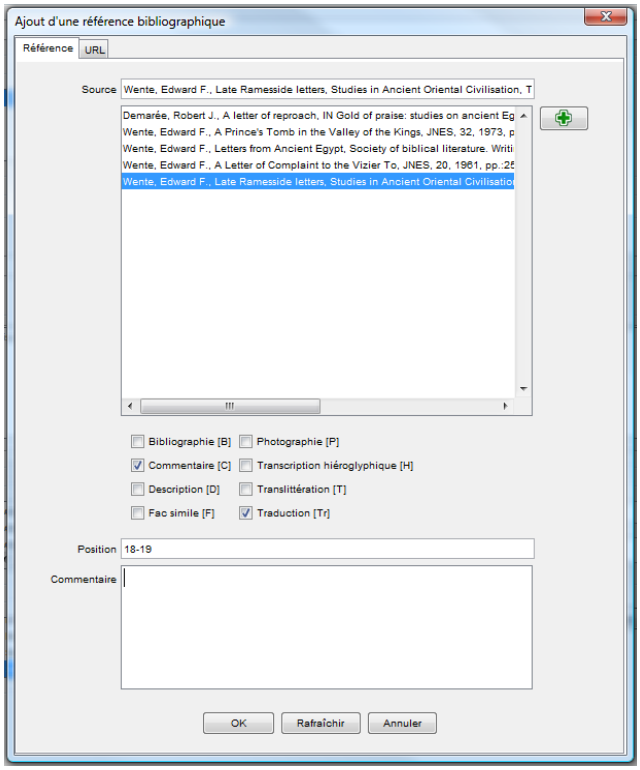

Figure 10. Linking a new reference

20. http://aaew.bbaw.de/tla/servlet/S04?f=h008.

Once a reference is linked to an entry in the database, a symbol identifies its type: [DIC] for dictionaries and lexica, [REF] for books and papers, and [URL] for websites[21] (see Fig. 11 with the lemma *ib*). For the sake of readability, references are listed according to these three major groups.

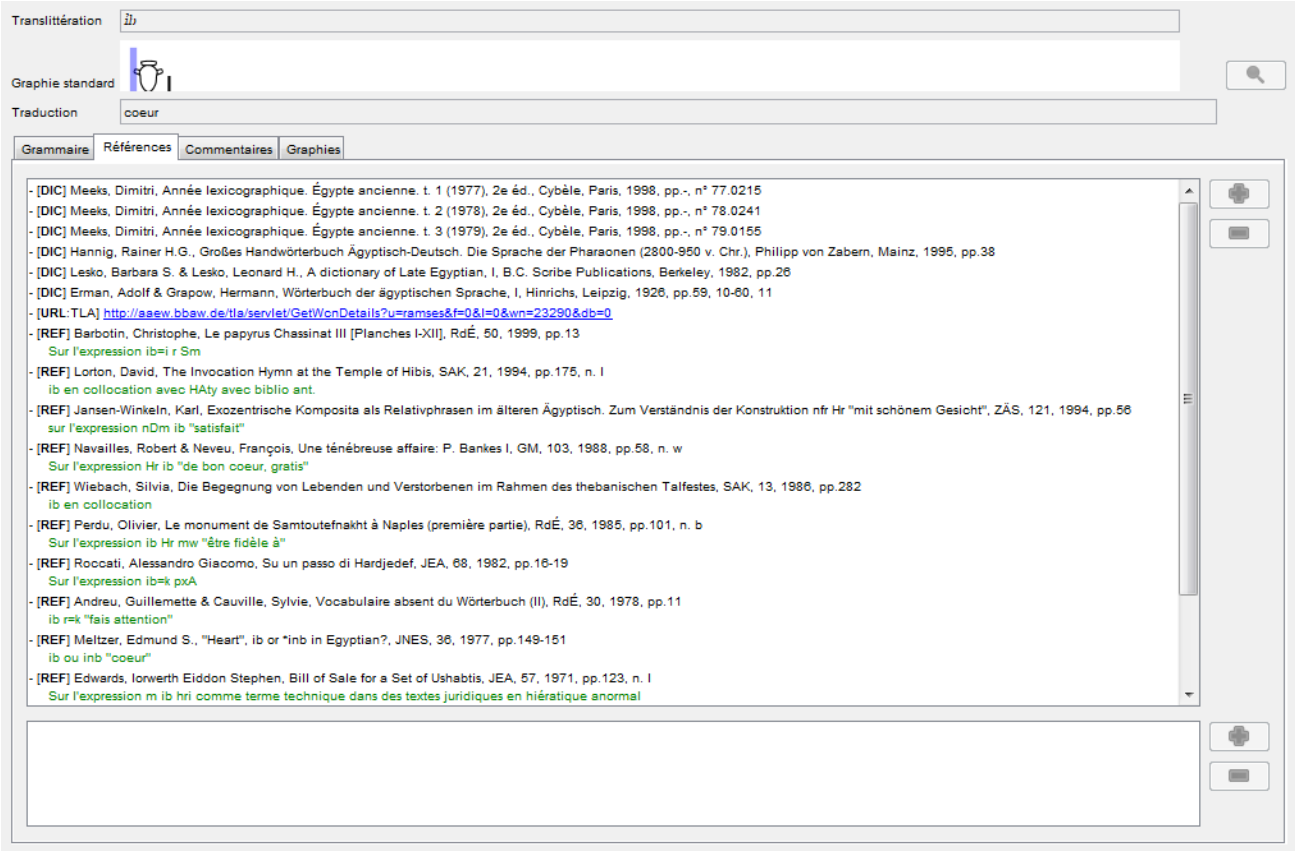

Figure 11. References linked to the lemma *ỉb* 'heart'.

3.2. *RamsesExport*

RamsesExport is an entirely new device that has been developed in order to meet to two urgent needs both for the team and the users:

(1) Double-checking the encoding. As stated earlier, Ramses is a manually annotated corpus; as such, the quality of the encoding is expected to be up to the highest standards, but at the same time human beings are notoriously fallible. Consequently, the annotations of each text are checked over twice in order to reach the highest possible degree of accuracy and consistency, which is hardly feasible while working on screen. A tool had to be developed in order to export all the data associated with a text in a printable format (`.pdf`).

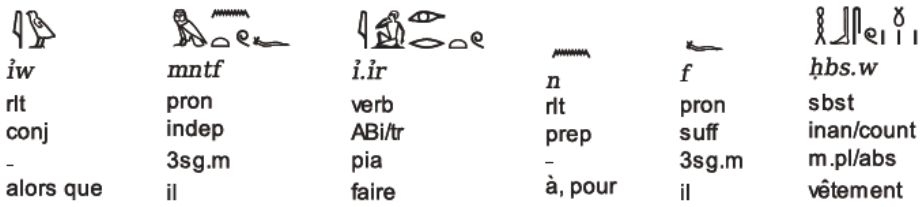

Figure 12. PDF export of the *Two Brothers* (one sentence on pD'Orbiney, l. 1,2)

21. It is worth mentioning that end-users can access online references directly from the Ramses interface via hyperlinks.

It is worth mentioning that besides entire texts, it is possible to export annotated data on specific sentences or sections of text, using in the latter case the reference system of the original document or of the edition:

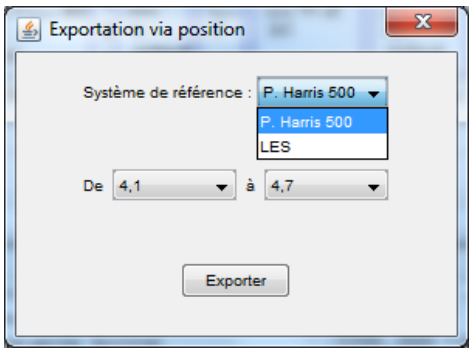

Figure 13. Exporting a section of a text using position

(2) Saving the results of complex searches. With the growing size of the database, some searches already produce a vast amount of results. In order to be able to deal with them properly when studying a given topic, it thus became crucial to enable end-users to export these results in a format convenient both for saving (the context, i.e. number of propositions before and after, of each result may be specified) and further treatments. It quickly became apparent that the HTML format was indeed well suited to such requirements (including the copy-pasting of hieroglyphs and glosses in a text document).

			(spelling ?)			{ — }	< spelling ? >			Ponct.	LES, 34,11
	bn	*iw*	*i*	(*r*)	*rḫ*	*rdi.t*	{ *s* }	< *f* >	*n*	*k*	
	gram	**gram**	**pron**	**gram**	**verb**	**verb**	**pron**	**pron**	**rlt**	**pron**	
	neg	**VerbP**	**suff**	**VerbP**	**2-lit/tr**	**anom/tr**	**suff**	**suff**	**prep**	**suff**	
	-	-	**1sg.m**	-	**inf/constr**	**inf/pron**	**3sg.f**	**3sg.m**	-	**2sg.m**	
	négation	iw (F3)	je	(r (F III))	apprendre à connaître	donner	{ elle (suffixe) }	< il >	à, pour	tu	
	2	**TRADUCTION : Je ne pourrai pas te le donner".**									

Figure 14. HTML export of one result of the search [lemma=*rḫ* + PoS=Verb]

As shown by Fig. 13-14, RamsesExport allows users to generate interlinear morphological glosses automatically. The types of data to be actually exported (hieroglyphs, morphological analysis, translation, data concerning textual criticism, etc.) can be selected before any export (in `.pdf` as well as in `.html`):

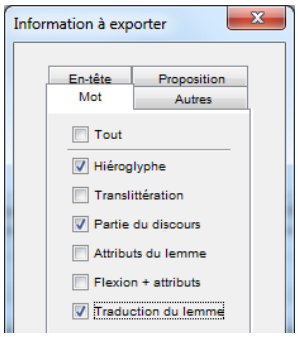

Figure 15. Selection of data to be exported with RamsesExport

4. THE SEARCH ENGINE

In this section, we present the latest version of Search Engine that has been the subject of considerable development and now allows (almost) any kind of query in the database.[22] Its nearly unlimited potential will assuredly be of paramount importance for the future studies in the fields of graphemics, morpho-syntax, onomastics, lexical semantics, and linguistic variation[23] in Late Egyptian.

We here focus on the significant features that have recently been added regarding (1) corpus selection and (2) search parameters.

(1) It is now possible to restrict a query to a part of the corpus using two filters. The first filter sets the time limits of the sub-corpus to be investigated. Users are offered the choice between a general selection by dynasties and more fine-grained selections by picking the name of a king (Fig. 16a):

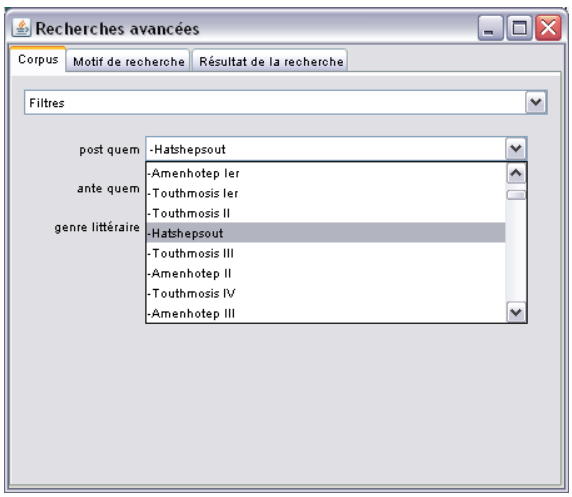

Figure 16a. Selection of a time frame

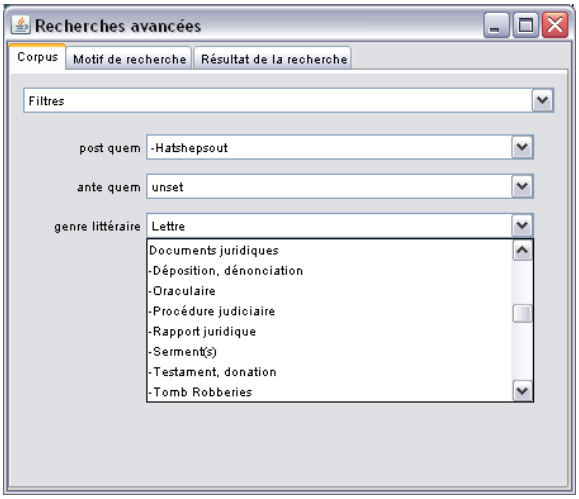

Figure 16b. Selection of a text genre

The second filter is related to the text genres. The user is presented with a dropdown list that contains the text genres identified in the corpus; the genres and sub-genres have been arranged in a hierarchical thesaurus so as to allow different degrees of precision in the queries (Fig. 16b). For instance, the category 'Administrative' is first subdivided into four classes: 'Private', 'Official', 'Lists' and 'Others'. To the class 'Official' belong two items: the so-called '*Journal de la Tombe*' and the 'Administrative Reports'. Accordingly, it is easy to select either broad genres, like administrative texts, or specialized sub-genres, like the documents belonging to the *Journal de la Tombe*.

As was already the case in previous versions of the Search Engine, the corpus may be defined manually (selection of texts in the list of annotated texts in the database) or by using the results of the last query as the corpus for a further query. The last option is a powerful tool for studying the lexicon, for it becomes possible to look e.g. for texts that contain pairs of closely related lexical items (cf. *infra*).

(2) Two general principles for building a query — already implemented in the previous versions of Search Engine — have been maintained:

- a query is based on (a sequence of) block-occurrence(s), corresponding roughly to a hieroglyphic spelling with all the annotations;
- a (theoretically) unlimited number of block-occurrences can be combined in a single query, either linearly or by using Boolean operators.

22. Regarding the diversity of possible searches in corpus linguistics, see Bilger 2000: 149-217.

23. See Rissanen (2008) for the use of corpora in historical linguistics (and especially in relation to a variationist approach to the study of language).

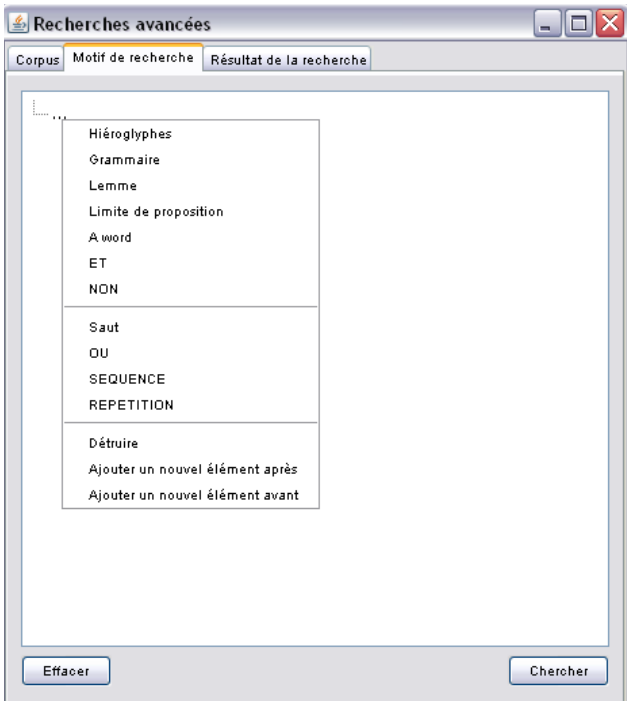

Figure 17. Options of a query

We first give a reminder of the options available for building a query on a single block-occurrence using the different levels of annotation in Ramses. Users can look for (a) a spelling, (b) a lemma or (c) a morphological analysis (and any combination thereof with the operator AND) (Fig. 17); the following simple queries exemplify different potentialities:

- a spelling that is not linked to any lemma or inflection (e.g. );[24]
- a single lemma (e.g. lemma = *rdi* "give");
- an inflection that is neither linked to a lemma nor to a spelling (e.g. all the occurrences of pseudo-participles);
- a lemma with a specific inflection (e.g. lemma *rdi* "give" + pseudo-participle, or *pr* "house" + plural);
- a lemma with a spelling (e.g. lemma *ḥ3b* "send" with the spelling );
- an inflection with a spelling (e.g. perfective passive participles with the ending )

Complex queries can be built by combining block-occurrences and using operators in order to define the relation between them. The following examples are valid queries in Ramses:

- Searching for the co-occurrence of two or more words in a proposition; e.g. *ḥ3b* "send" and *šʿ.t* "letter". It is possible to look for contiguous words or to allow for some space between two words (using the SKIP operator). For instance, *ḥ3b* "send" + max. 3 words + *šʿ.t* 'letter' is a possible request.
- The same request can be made with some additional morphological precisions. For instance, one can limit the query to *ḥ3b* in the imperative. As seen before, this is achieved by using the operator AND that allows for combining different criteria on the same block:

24. The possibility of looking for a sequence of signs within a word is particularly useful for studying graphemic classifiers in Ancient Egyptian. One could, for instance, search for the sequence  , which is almost exclusively found as a combined classifier. This option is also useful for filling in lacunae when editing a new text.

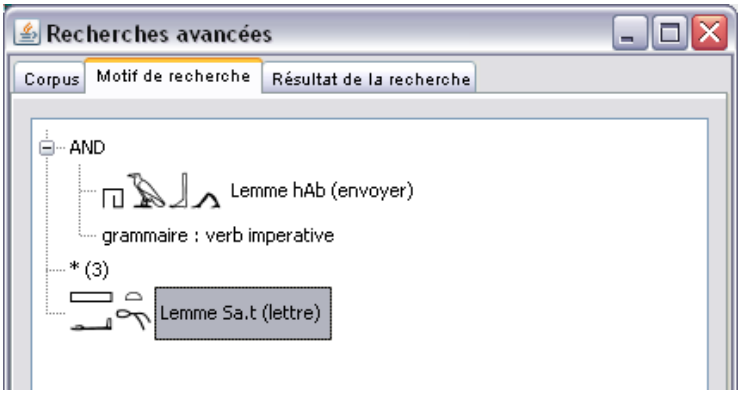

Figure 18. Using the AND and SKIP (*) operator

The next figure shows how one of the results is highlighted within the Ramses interface:

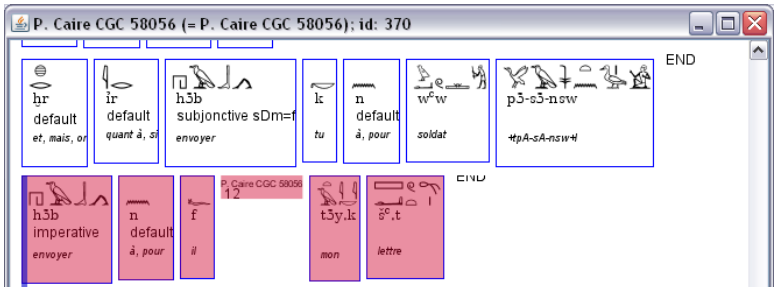

Figure 19. Display of one of the results of the query displayed in Fig. 18

– By combining the Boolean operator OR with the possibility to select the result of a previous query as corpus of a new search, one can study the collocation of lexical or grammatical synonyms in the corpus. The next figure shows the occurrences of *ḥ3ty* and *ib* when appearing in the same texts. The procedure is as follows: first, look for *ḥ3ty* (or *ib*) in the whole database; second, select the result as the corpus for the following request; third, look for *ib* (if *ḥ3ty* was chosen in the first request).

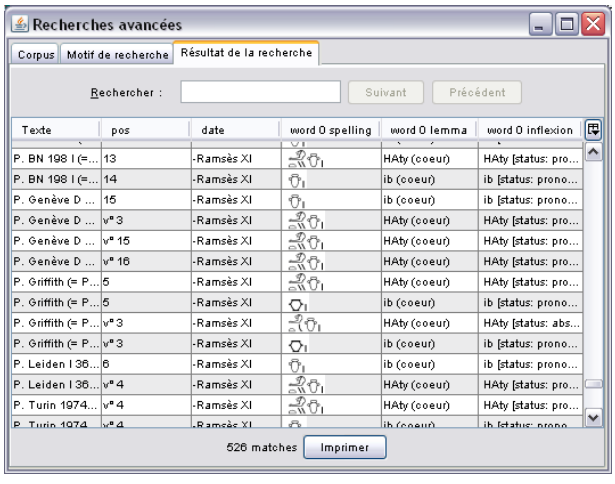

Figure 20. *ḥ3ty* and *ib* occurring in the same texts

This procedure can also be used to investigate grammatical facts as, for instance, variants of a grammatical pattern, i.e. pairs like *n3-n* vs. *n3* (ART.PL), *-sn* vs. *-w* (3PL pron.), *ḥn˓ ntf sḏm* vs. *mtw.f sḏm*, *i.sḏm.t.f* vs. *i.ir.t.f sḏm*, etc. This of course also applies for variations at the graphemic level.

– A new operator that has been added is REPEAT. It enables users to spot contiguous repetitions of lemmata, inflections or graphemes. The number of repetitions can be fixed (with a

minimum and a maximum). Consequently, it is possible to test whether there are examples of three adjectives in a row (there are!), or if a definite article can be repeated (in order to study cases of dittography). Detecting a repetition of a phoneme is possible. By combining REPEAT with the operator SEQUENCE, also a newcomer in Ramses, it is possible to build sophisticated queries. For instance, one can look for multiple predicates in a conjugation pattern. In the database, there are a few occurrences of two coordinated [(*ḥr*) + INF.] in the sequential patterns:

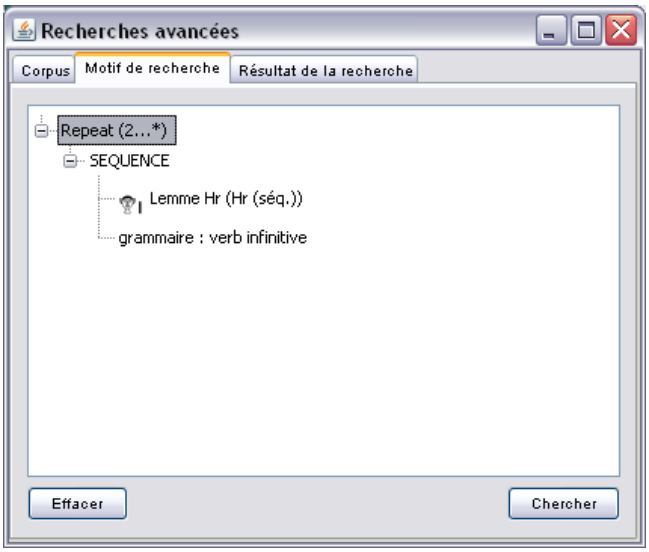

Figure 21. Using the operators REPEAT and SEQUENCE in combination

– The use one can make of these two operators seems to be limited only by imagination. We realized, for example, that they can be used to detect some particular uses of the classifier G7, that sometimes plays the role of a cohesive marker as in the following example:

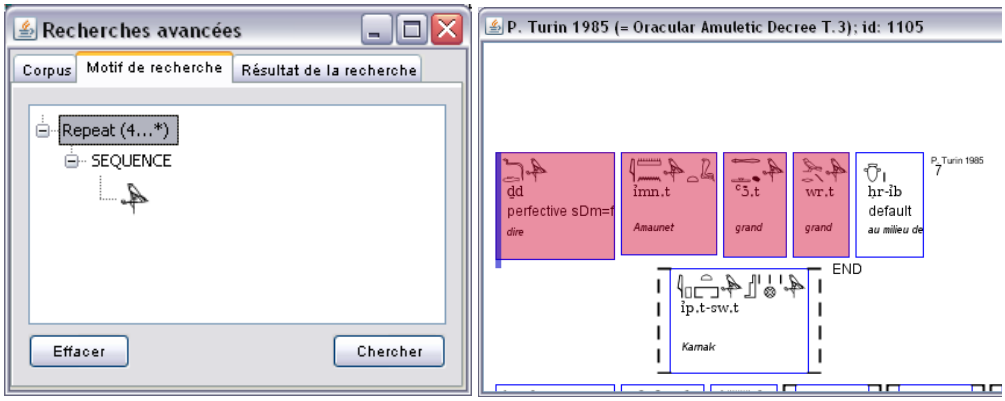

Figure 22. The classifier G7 as a semantic cohesive marker

– By default, the requests were first limited to a single proposition. It is now possible to cross this limit using the operator PROPOSITION END. This considerably extends the possibilities of the search engine. For instance, one can look for a combination of verbal patterns: any verb in imperative + any verb in the conjunctive is now a possible query in Ramses. The following figure illustrates one of the results

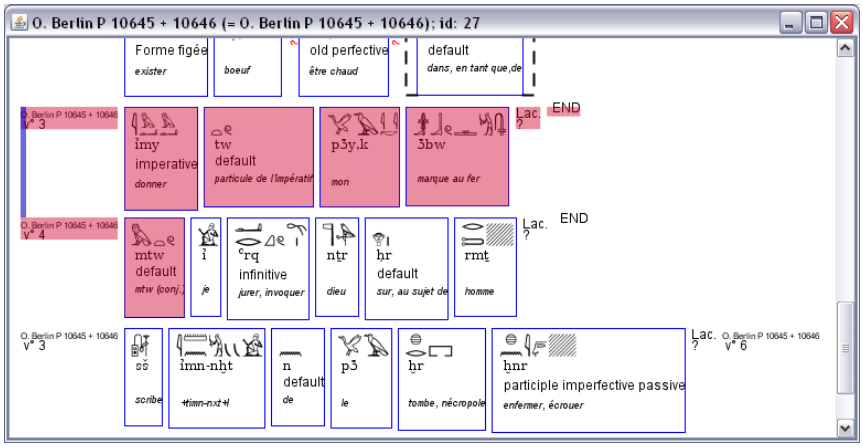

Figure 23. Looking for imperatives followed by conjunctives

Another possible approach to the same case study would be to look for any kind of verbal pattern that precedes a conjunctive. This kind of request is not ideal, but satisfactory for the time being while the syntactic analysis is still under construction (see Polis & Rosmorduc in the current volume). This query thus produces too many results because the verbal pattern in the first proposition is not always syntactically on the same level as the conjunctive in the next proposition. The following figure shows the result of such a request sorted out according to the verbal inflections of the first proposition:

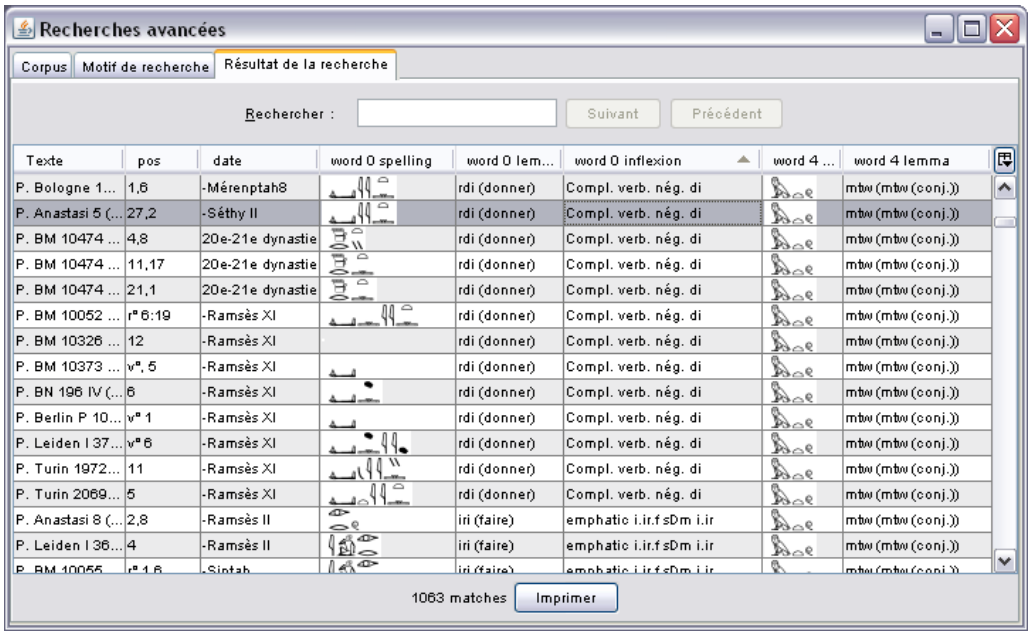

Figure 24. Looking for patterns occurring before conjunctives

5. SHORT- AND LONG-TERM PERSPECTIVES

Before termination of the first phase of the project in October 2013 (end of the 'ARC' founding, see §2.1), 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 (see §2.2; with a particular focus on the non-narrative literary texts, on the judicial documents, on the texts of the Third Intermediate Period and on the texts written in so-called "abnormal hieratic").

(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 (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? etc.).

(3) Implementation of context sensitive semi-automatic part-of-speech tagger and syntactic parser (topic of Benjamin Martin Leon's PhD thesis) in order to facilitate the annotation of new texts in the future and ensure *a priori* the coherence of the annotations.

(4) Implementation of new search functions (especially at the syntactic level) and development of additional sorting facilities (e.g. data sorted not only according to time, but also according to the place of origin of the documents, the writing system, etc.).

(5) Development of a Web application that would give the community of Egyptologists and linguists access to the whole range of the Ramses data. We plan to publish the sub-corpora online in sequence directly after final approval of the team. In order to allow the end-users to contribute to the enrichment of the corpus, a wiki-like device will be added in order to allow suggestions regarding the hieroglyphic readings, the addition or emendation of annotations, etc.

Long-term projects include:

- The standardization of the thesauri on which the Ramses annotation scheme is based, including e.g. the matching of the actual geographical thesaurus with the *Multilingual Egyptological Thesaurus*,[25] and the matching of the idiosyncratic tagset for morphological annotation with emergent *de facto* standards (like EAGLES,[26] Multext, etc.).
- The completion of the syntactic annotation of the corpus and the addition of a semantic level of annotation (with word-sense disambiguation).
- The continuation of existing (and development of new) collaborations, e.g. with TXM (see §2.1) concerning statistic tools, with the *Thesaurus Linguae Aegyptiae* (see Dils & Feder in the current volume) in the field of Egyptian lexicography, with the Deir el-Medina Database (see n. 18) regarding the metadata on Late Egyptian texts, etc.
- The extension of Ramses functionalities in order to be able to deal with earlier and later stages of the language (down to Coptic).

BIBLIOGRAPHY

BILGER, Mireille (ed.). 2000. *Corpus : Méthodologie et application linguistique*, Paris, Honoré Champion (= Bibliothèque de l'INaLF, Les français parlés – Textes et études 3).

CLARIDGE, Claudia. 2008. Historical corpora, in: Hanke LÜDELING & Merja KYTÖ (eds.), *Corpus linguistics. An International Handbook*, Vol. 1, Berlin-New York, de Gruyter Mouton (= HKS 29.1), p. 242-259.

DAVID, Arlette. 2006. *Syntactic and Lexico-Semantic Aspects of the Legal Register in Ramesside Royal Decrees*, Wiesbaden, Harrassowitz (= Göttinger Orientforschungen IV/38).

DILS, Peter & Frank FEDER. Current volume. The *Thesaurus Linguae Aegyptiae.* Review and perspectives.

GARSIDE, Roger, Geoffrey N. LEECH & Tony MCENERY (eds.). 1997. *Corpus annotation: Linguistic information from computer text corpora*, London, Longman.

HONNAY, Anne-Claude & Stéphane POLIS. 2011. Projet Ramsès. Manuel d'encodage, in: http://www.egypto.ulg.ac.be/docs/Ramses_Manuel_2011.pdf

HUNSTON, Susan. 2002. *Corpora in Applied Linguistics*, Cambridge, Cambridge University Press.

LEECH, Geoffrey. 1993. Corpus annotation schemes, in: *Literary and Linguistic Computing* 8/4, p. 275-281.

25. See van der Plas 1996.

26. See Leech *et al.* (1996) who provide a language-neutral "intermediate tag set", using a numeric coding for each feature.

—. 2005. Adding linguistic annotation, in: M. WYNNE (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, Oxford, Oxbow Books, p. 17-29

LEECH, Geoffrey, Ros BARNETT & Peter KAHLER (eds.). 1996. EAGLES Final Report and Guidelines for the Syntactic Annotation of Corpora, EAGLES Document EAG-TCWG-SASG/1.5, Lancaster University.

LEM = Alan H. GARDINER. 1937. *Late Egyptian Miscellanies*, Brussels (= Bibliotheca Aegyptiaca 7).

LES = Alan H. GARDINER. 1932. *Late Egyptian Stories*, Brussels (= Bibliotheca Aegyptiaca 1).

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

LRL = Jaroslav ČERNÝ. 1939. *Late Ramesside Letters*, Brussels (= Bibliotheca Aegyptiaca 9).

LRLC = Jac J. JANSSEN. 1991. *Late Ramesside Letters and Communication*, London (= Hieratic Papyri in the British Museum 6).

MAZZIOTTA, Nicolas. 2010a. Logiciel NotaBene pour l'annotation linguistique. Annotations et conceptualisations multiples, in: *Recherches qualitatives. Hors série "Les actes"* 9, p. 83-94.

—. 2010b. Build the *Syntactic Reference Corpus of Medieval French* Using *NotaBene RDF Annotation Tool*, in: *Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV)*, Uppsala, Association for Computational Linguistics (ACL), p. 142-146

VAN DER PLAS, Dirk (ed.). 1996. *Multilingual Egyptological Thesaurus*, Utrecht-Paris, Centre for Computer-aided Egyptological Research (= Publications Interuniversitaires de Recherches Égyptologiques Informatisées 11).

POLIS, Stéphane. 2006. Le projet Ramsès, in: Jean Winand, Un siècle d'Égyptologie à l'Université de Liège, in: Eugène WARMENBOL (ed.), *La caravane du Caire. L'Égypte sur d'autres rives*, Louvain-la-Neuve, Versant Sud, p. 180.

POLIS, Stéphane & Serge ROSMORDUC. Current volume. Building a construction-based Treebank of Late Egyptian. The syntactic layer in Ramses.

RAD = Alan H. GARDINER. 1948. *Ramesside Administrative Documents*, London.

RISSANEN, Matti. 2008. Corpus linguistics and historical linguistics, in: Hanke LÜDELING & Merja KYTÖ (eds.), *Corpus linguistics. An International Handbook*, Vol. 1, Berlin-New York, de Gruyter Mouton (= HKS 29.1), p. 53-68.

ROSMORDUC, Serge, Stéphane POLIS & Jean WINAND. 2009. Ramses. A new research tool in philology and linguistics, in: Nigel STRUDWICK (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*, New Jersey, Gorgias Press (= Bible in Technology 2), p. 133-142.

TR = T. Eric PEET. 1930. *The Great Tomb-Robberies of the Twentieth Egyptian Dynasty. Being a Critical Study, with Translations and Commentaries, of the Papyri in which these are Recorded*, 2 vol., Oxford.

WINAND, Jean, Stéphane POLIS & Serge ROSMORDUC. In print. Ramses. An annotated corpus of Late Egyptian, in: Panagiotis Kousoulis & Nikolaos Lazaridis (eds.), *Proceedings of the Tenth International Congress of Egyptologists. University of the Aegean, Rhodes, 22-29 May 2008*, Leuven, Peeters (Orientalia Lovaniensia Analecta), 10 p.

Abstracts

Peter Dils & Frank Feder, The *Thesaurus Linguae Aegyptiae*. Review and Perspectives

The *Thesaurus Linguae Aegyptiae* (*TLA*) represents today the largest available database of Egyptian texts and, moreover, it is worldwide accessible on the Internet with free access. It combines a text corpus of Egyptian texts from nearly all periods of Egyptian history with an electronic lexicon. Both are linked to each other and are regularly updated. The TLA provides also access to the digitalized material on which the edition of the *Wörterbuch der aegyptischen Sprache* was based (slip archive). The text corpus and the lexicon can be searched in a number of ways and for different purposes; tools for statistical analysis are provided as well. As the *TLA* is a dynamically developing database system the text corpus and the lexicon will further be expanded, especially by adding the still lacking Coptic material of the Egyptian language, and by improving the research tools gradually.

Stéphane Polis, Anne-Claude Honnay & Jean Winand, Building an Annotated Corpus of Late Egyptian. The Ramses Project: Review and Perspectives

This paper reviews the experience of the Ramses Project in constructing a richly annotated corpus of Late Egyptian that consists of 300 000 words in 2011 (and is expected to grow up to more than 1 million words in coming years). During the first five years of the project, this corpus has been encoded in hieroglyphic script, translated in French or English and received annotations for part-of-speech information, lemmatization, and morphological analysis. The methodology and working tools that have been developed in order to build this corpus are here discussed and future developments are presented.

Stéphane Polis & Serge Rosmorduc, Building a Construction-Based Treebank of Late Egyptian. The Syntactic Layer in Ramses

This paper reports on the construction-based Treebank currently under development in the framework of the Ramses Project, which aims at building a multifaceted annotated corpus of Late Egyptian texts. We describe the specifications that have been implemented and we introduce the syntactic formalism and the related representation format that are used for the syntactic annotation. Furthermore, the annotation scheme is discussed with particular attention paid to its evolutionary nature. Finally, we explain the methods as well as the annotating tool, called *SyntaxEditor*; we conclude by

addressing the question of forthcoming developments, especially the search engine and a context-sensitive parser.

Stéphanie GOHY, Benjamin MARTIN LEON & Stéphane POLIS, Automated Text Categorization in a Dead Language. The Detection of Genres in Late Egyptian

This paper is a first step in applying machine learning methods typical of Automated Text Categorization (ATC) for Automatic Genre Identification (AGI) in Late Egyptian, a language written in either hieroglyphic or hieratic scripts that is found in documents from Ancient Egypt dating from ca. 1350-700 BCE. The study is divided into three parts. After a general introduction on AGI (§1), we introduce the levels of annotation that are integrated in the Ramses corpus and can be used when performing AGI on Late Egyptian (§2). In the following section (§3) we offer a brief survey of the types of features that have been discussed in the literature on AGI, before proceeding with three case studies where we apply supervised machine learning methods — namely the naïve Bayes classifier (§4.1), the Support Vector Machine (§4.2), and the Segment and Combine approach (§4.3) — to a selection of texts in the corpus. Their respective performances are tested using lexical, part-of-speech and inflectional features.

Mark-Jan NEDERHOF, Flexible Use of Text Annotations and Distance Learning

In this paper, we discuss a framework that allows independently created annotations of texts to be combined and presented as one unified interlinear format. Applications for distance learning are also considered. As proof-of-concept, we present PhilologEg, a tool that can be used to study an Ancient Egyptian hieroglyphic text in combination with any number of translations and grammatical annotations. The tool is a fully integrated system that runs on all major platforms.

Roberto GOZZOLI, Hieroglyphic Text Processors, Manuel de Codage, Unicode, and Lexicography

This paper gives an overview of the different software available to scholars working in the field of Egyptian language, with a special focus on hieroglyphic typesetting, Unicode and lexicographical databases that systematically encodes hieroglyphs. Various problems with the *Manuel de Codage* are discussed, as well as the need for a more active interaction between computers and Egyptology. A proposal for Egyptological software is given at the end of the paper.

Mark-Jan NEDERHOF, The Manuel de Codage Encoding of Hieroglyphs Impedes Development of Corpora

In this paper, we discuss the encoding of hieroglyphic text and argue that the set of requirements for an encoding scheme depend on the intended application. Our main claim is that if this application is the development of text corpora with long lifespans and diversity of use, then encoding schemes within the tradition of the Manuel de Codage are unsuitable.

Vincent EUVERTE & Christian ROY, Hieroglyphic Text Corpus. Towards Standardization

Sharing the heritage of Ancient Egyptian written production means facing numerous technical challenges. The goal of this paper is to build a preliminary inventory of these challenges and to propose some possible solutions. After a quick overview of the topics that are possible candidate to an international standardization, the paper focuses on two aspects. (1) The 'Multilingual Egyptological Thesaurus' (MET), initiated in 1996 by Dirk van der Plas, has not changed since 2003. It could be updated and expanded with minimal effort under the coordination of an official body such as the Center for Documentation of Cultural and Natural Heritage (CULTNAT). (2) The 'Manuel de Codage' (MdC) has not benefited from developments in computer science since the third edition was

published under the *Informatique & Égyptologie* mandate in 1988. Over time, each hieroglyphic software program has developed its own specific syntax to satisfy emerging needs, making it difficult for users to share ancient Egyptian texts. For these two topics, we will suggest a plan for improvement based on the Rosette Project's experience, though the input of the Egyptologists' community at large is appreciated to refine various concepts and identify the best route forward.

Christian MADER, Bernhard HASLHOFER & Niko POPITSCH, The MEKETREpository. A Collaborative Web Database for Middle Kingdom Scene Descriptions

Whilst representations, iconography and the development of scenes in private and royal tombs from the Old Kingdom have been studied extensively in the past, comparable research of Middle Kingdom (MK) representations and scene details is still underrepresented. The MEKETRE research project aims at closing this gap by systematic research of MK representations. In the course of this project, an online digital repository (the MEKETREpository) is being built that enables researchers to describe and annotate MK two-dimensional art at various levels of detail using images, free text, and controlled vocabularies. It also enables the collaborative development of semantic vocabularies for the description of these data. The MEKETREpository will publish the resulting data and vocabularies as Linked Data on the Web by utilizing Semantic Web technologies to enable their integration into other Linked Data sets such as DBpedia, Freebase or LIBRIS. The collected data is described using standardized and specialized vocabularies allowing for easy integration into existing databases and search engines. For the long-term preservation of the data, the MEKETREpository will make use of the University of Vienna's digital asset management system PHAIDRA. At its final stage the MEKETREpository will supply a platform that exposes collaboratively created, continuously evolving, and publicly available information about the MK on the Web.

Nathalie PRÉVÔT, The Digital Puzzle of the *talatat* from Karnak. A Tool for the Three-Dimensional Reconstruction of Theban Buildings from the Reign of Amenhotep IV

The revival of studies on the Atonist temples of Karnak (program of the French National Research Agency ATON-3D – ANR-08-BLAN-0202-01) required the implementation of an Information System dedicated to the Theban *talatat* that would also be accessible to the scientific community. This IS is associated with software which helps to reassemble the fragmented reliefs (a digital interactive puzzle), constituting a real tool for researchers and providing the knowledge needed to produce and validate hypotheses about the structures and dimensions of the buildings. The database is then enriched with images of the temple's extrapolated decoration, which involves 3D modelling of these extrapolations. *Talatat* indexing was based on the Multilingual Egyptian Thesaurus conventions regarding "passport" data, including iconographic description using descriptive operators called *unicos*. In the spirit of the international movement in favour of open access to scientific data, the *talatat* metadata and images are accessible online to researchers working on the proto-Amarna or Amarna periods. The *talatat* metadata is published using RDFa data model mapping for embedding RDF triples within the XHTML of our web pages, which can be extracted by compliant user agents. This corpus is stored in a secured warehouse with strong human and digital infrastructure for preservation of the images and of their metadata.

Carlos GRACIA ZAMACONA, A Database for the Coffin Texts

This article describes a database for the Coffin Texts. It was first conceived as a semantic study of verbs of motion, and for this reason many of its files are linguistically focused. Nevertheless, it may be useful for other kinds of studies, because the software employed allows integration of new files as well as modification of old ones. This is the ultimate aim of such a database: a tool appropriate for all kinds

of research on this corpus. Specific features of this corpus are discussed first, followed by the database conception and structure, and finally its use, results and developments.

Azza EZZAT, The Digital Library of Inscriptions and Calligraphies

The Digital Library of Inscriptions aims at recording all inscriptions on ancient Egyptian buildings and monuments throughout the ages. These inscriptions are digitally displayed for the user, including a brief description and pictures of the inscriptions. The languages included in the Digital Library are Ancient Egyptian, Arabic, Turkish, Persian and Greek languages. Moreover, there are inscriptions bearing Thamodic, Musnad, and Nabatean scripts.

Yannis GOURDON, The *AGÉA* Database Project.
Anthroponymes et Généalogies de l'Égypte Ancienne

Since the 30s, our understanding of the ancient Egyptian personal names has been dependent on Ranke's *Personennamen*. But, because the data and its philological and sociological analysis are based on the knowledge available in the first half of the 20th century, the *PN* requires a complete revision that takes into account recent developments on the subject. Launched in 2008 at the IFAO, the *AGÉA* database project aims, eventually, to create a systematic directory of personal names for every period of the Pharaonic history, completing and modernizing Ranke's work. As a tool facilitating more efficient analysis and a better interpretation of data, *AGÉA* will focus, in its first development, on the Old Kingdom.