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
9 782875 620163

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

1. *Annotated corpora of Ancient Egyptian texts*

2. *Hieroglyphic encoding*

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

Texts, Languages & Information Technology in Egyptology

Introduction

Stéphane POLIS
F.R.S.-FNRS – Université de Liège

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, they discuss the most promising avenues for future research, developments and implementation, and they suggest solutions to longstanding issues in the field.

Two general trends characterize the projects laid out here: the will to be available online for 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/or exchanged without loss of information.

1. ANNOTATED CORPORA OF ANCIENT EGYPTIAN TEXTS

The volume opens with papers on two large-scale collective projects of annotated corpora in Ancient Egyptian. The first is the *Thesaurus Linguae Aegyptiae*, a major achievement of recent decades that is now part of every Egyptologists' daily life: it represents the largest database of Egyptian texts (with over 900 000 tokens) and it is freely available online. Peter DILS & Frank FEDER introduce the database structure and they outline the texts that have been included in the *TLA* corpus so far. Furthermore, they provide an overview of the searching, sorting and counting facilities that are accessible to anyone on the Internet and present the tool that has been developed for handling hieroglyphic spellings as well as the promising pictorial dictionary and image database that are being appended to the existing material.

The second project is the much younger *Ramses Project*. As stressed in the paper by Stéphane POLIS, Anne-Claude HONNAY & Jean WINAND, this project is more limited in terms of chronological scope, since it focuses on the corpus of Late Egyptian texts broadly speaking (from the 18th dynasty down to the Third Intermediate Period). The limited size of the corpus (c. 300 000 tokens as of late 2011) has the advantage of allowing for the systematic encoding of normalized hieroglyphic spellings (c. 45 000 spellings) as well as for detailed morphological analysis. In the near future, the corpus will also include a layer of syntactic analysis. In a separate contribution, Stéphane POLIS & Serge ROSMORDUC report on the construction-based Treebank currently under development for Ramses, with an introduction to the syntactic formalism and representation format that are used for this syntactic annotation.

Alongside the new search facilities that are offered by such annotated corpora, an entirely new field of research can now be investigated in Ancient Egyptian, namely that of Natural Language Processing (NLP), with the development of tools such as taggers and parsers. The interest of working on tools of this kind should not be underrated: besides the evident advantages in terms of speed of annotation, the corpora under construction could benefit from these techniques so as to enhance both consistency and accuracy of annotation. In this framework, Stéphanie GOHY, Benjamin MARTIN LEON & Stéphane POLIS describe in their contribution a pilot study on Automated Text Classification: three Machine Learning methods are applied to Late Egyptian texts in order to identify automatically the genre to which they belong. The goal of this inquiry is twofold: on a linguistic level, it works as a heuristic tool for evaluating the types of linguistic features that are characteristic of each genre, while on a more practical level, automatic genre identification is known to enhance the performance of taggers and parsers that can adapt to the specific norms of the genres.

Both the *TLA* and *Ramses* are supported institutionally and agreements have been made within each team of scholars about the levels of annotation, the related conventions, and the methods of handling problematic cases. Mark-Jan NEDERHOF suggests, in a first contribution, a promising way of dealing with the creation of less centralized forms of multilevel annotated corpora, with minimal requirements in terms of file format and convention agreements. The idea is to develop sophisticated software that can process text annotations coming from various sources and to render them in a uniform interlinear format. The author shows — thanks to the proof-of-concept *PhilologEg* — that the required tool can be realized and used to study Ancient Egyptian hieroglyphic texts in combination with any number of translations and grammatical annotations.

2. HIEROGLYPHIC ENCODING

The Ancient Egyptian hieroglyphic (and to a large extant hieratic) writing system has a number of properties — e.g. high level of iconicity, complex arrangement of the signs, use of graphemic classifiers — that set it apart from most of the world's writing systems. Hence the issues linked to its encoding are not trivial: how do we distinguish characters from glyphs (I refer here, *inter alia*, to the process that led in 2010 to the addition of 1071 hieroglyphic signs to Unicode 5.2); what is the level of precision that is needed in the rendering of any individual sign (depending on the field of use, e.g. palaeography, grammar, etc.); how precise must be the relative positioning of signs?

The so-called "Manuel de Codage" (1988) was the first answer by Egyptologists (*Informatique & Égyptologie* 2) to the challenge of defining a scheme for encoding normalized hieroglyphs. Over the years, however, this "standard" has been interpreted in various ways and received several sorts of additions in the hieroglyphic editing systems that were successively developed. As stressed by Roberto GOZZOLI in his overview of the tools that exist for hieroglyphic typesetting (also considering Unicode and lexicographical databases that encodes hieroglyphs), the versatility of the encoding scheme progressively led to the present — undesirable — situation where the lack of interoperability (and the related reduplication of work) is the norm.

This state of affairs is especially problematic, as Mark-Jan NEDERHOF argues in his second paper, for the development of hieroglyphic text corpora with long lifespans and a diversity of research applications. The author insists that such corpora should rely on an encoding scheme that (1) is stable, (2) has a high expressive power while remaining simple, (3) has operators with precise meaning, (4) is font-independent, and (5) is flexible in terms of formatting. Stepping out of the publication oriented (pseudo-facsimile) uses of the Manuel de Codage, he presents the principles of a new encoding scheme — the Revised Encoding Scheme (RES, first introduced in 2002) — that has been designed in order to meet these five requirements.

This need for standardization is stressed again in Vincent EUVERTE and Christian ROY's contribution.[1] Instead of a new encoding scheme, however, they suggest further developments for the Manuel de Codage that would lead to the inclusion of new functionalities, while stressing the need for an updated syntax. The principles argued for are illustrated based on the experience of the Rosetta project.

It appears that while the suggested solutions may be different, the acknowledgment remains identical: a revision of the Manuel de Codage is greatly desired. It will be up to the Computer Working Group to make suggestions in this direction to the International Association of Egyptologists in the near future.

3. DATABASES FOR ART HISTORY, TEXTS AND PROSOPOGRAPHY

The third part of this volume is dedicated to the presentation of databases — most of which are already accessible online — that have been developed in the field of history of art, textual material and prosopography:

- Christian MADER, Bernhard HASLHOFER & Niko POPITSCH present the *MEKETREpository*, a collaborative Web database that enables scholars to describe and annotate Middle Kingdom two-dimensional art at various levels of detail using images, free text, and controlled vocabularies. This database is part of the MEKETRE research project — that aims at researching the Middle Kingdom representations in a systematic fashion — and conforms to the latest developments in terms of standards and Web technologies. The repository is now freely available online and will undoubtedly be a reference for any forthcoming project in the field.
- In her paper, Nathalie PRÉVÔT describes a software solution (*Archeogrid*) that allows reassembling the fragmented reliefs of the Atonist temples from Karnak that are found on *talatat*, a digital interactive puzzle. This tool makes use of metadata on the *talatat* (RDFa data model mapping) and helps to produce and validate hypotheses about the structures and dimensions of the buildings in the framework of the ATON-3D project.
- Carlos GRACIA ZAMACONA gives an overview of his database of the *Coffin Texts*. He first conceived it in order to facilitate the study of the verbs of motion in this specific corpus. However, the ultimate goal of the database is to serve as a tool for all kinds of research on the *Coffin Texts*, which would require the completion of the current encoding work and the addition of other types of data by a larger team of scholars.
- Azza EZZAT offers a general presentation of *The Digital Library of Inscriptions and Calligraphies*, an ambitious project that aims at recording eventually all inscriptions on ancient Egyptian buildings and monuments throughout the ages. The Web interface gives nowadays access to many types of artifacts bearing inscriptions in Ancient Egyptian (with a brief description and pictures of the inscriptions). Alongside Ancient Egyptian, other languages attested in Egypt throughout the ages (such as Arabic, Turkish, Persian and Greek) are considered.
- In his paper, Yannis GOURDON introduces the *AGÉA* database (*Anthroponyms and Genealogy of Ancient Egypt*). This project began in 2008 at the Institut Français d'Archéologie Orientale with the aim of creating a systematic directory of personal names for every period of the Pharaonic history, completing and modernizing the previous standard work by Hermann Ranke. In its first phase, *AGÉA* focuses on data of the Old Kingdom. The present paper systematically surveys the database structure and design. It is available online in a beta version since late 2011.

[1] Another candidate for international standardization is the 'Multilingual Egyptological Thesaurus' (MET) that could be updated and expanded with minimal effort, as the authors suggest, under the coordination of an official body such as the Center for Documentation of Cultural and Natural Heritage (CULTNAT).

This volume closes with a paper by Eugene Cruz-Uribe on computer and journal publishing. The author discusses the pros and cons of using new technologies in journal publishing. Both as an editor and Egyptologist, his position is that it will be more and more difficult to support hard copy journal publishing and that within a reasonable timeframe of 15 years, all journals should have moved online. At the same time however, web technologies for publication should not be endorsed without a clear sense of the implications that this shift will have on our publication methods and practices. In this respect, he stresses the need for a standard hieroglyphic encoding scheme and insists on the development of related rendering tools for printed material (cf. §2). Furthermore, all journals — he argues — should plan to convert entirely to online format and use this opportunity to redefine their goals and favorite topics among the large fields of research that Egyptology encompasses.

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 descripttion 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.