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

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.

The *Thesaurus Linguae Aegyptiae*

Review and Perspectives

Peter DILS & Frank FEDER

Leipzig – Berlin

1. INTRODUCTION

The research project *Altägyptisches Wörterbuch*, approved by the Union of the German Academies of Sciences in 1992, was established at the Berlin-Brandenburg Academy of Sciences and Humanities (*BBAW*). The aim of the project was to develop a comprehensive annotated electronic corpus of ancient Egyptian texts in a digital database, which would constitute a powerful research tool in its own right while serving as a source for compiling a dictionary in the future.

Work started at a time when there was little experience with large electronic databases of texts — and in particular, of texts composed in dead languages as well as in non-Latin scripts; when personal computers were not very powerful; and when the format for making such a database accessible was open to suggestions.

The project has advanced considerably beyond its initial stage and can justly be called a technical, material, and methodological success. On the technological side, a computer program, the *Berliner Texterschließungssystem* (*BTS*: Berlin Text Encoding System), was developed for encoding ancient Egyptian texts in transcription and translation, as well as for providing them with lexical and grammatical annotations. The project soon realised the potential of the *World Wide Web* and solved the problem of accessibility by publishing the database on the Internet as the *Thesaurus Linguae Aegyptiae* (*TLA*).[1] The material included in the database comprises hundreds of texts with more than 900 000 text words at last count. In addition, the 1 700 000 text words[2] recorded on paper slips, which served as the basis for compiling Erman and Grapow's *Wörterbuch der aegyptischen Sprache,* available online since 1999 as the *Digitalisiertes Zettelarchiv*, are now integrated into the *TLA*. As far as methodology is concerned, the *TLA* advanced beyond the standard search and sort functions of a database to implement a number of statistical procedures adopted from the field of corpus linguistics.

The progressive nature of the *TLA* attracted other research institutions both outside and within Germany, who joined the project as cooperating partners. Nowadays, the *Thesaurus Linguae Aegyptiae* is the largest database of ancient Egyptian texts worldwide and, moreover, the only publicly accessible database that facilitates statistical analyses.

[1] http://aaew.bbaw.de/tla/index.html.

[2] 1 200 000 text words in context in alphabetical order; 1 700 000 when the *Sonderverzettlung* of prepositions, of personal and royal names, etc. is included.

2. THE DATABASE STRUCTURE

The project uses a relational database with two main modules:

(1) a continually expanding lexical database of all known Egyptian lexemes — a kind of electronic lexicon, with basic translation of each lexeme, assignment of it to a particular class of words, and bibliographical references;

(2) a continually expanding text database in which each individual text is divided into sentences and phrases with the text words linked to the lexical database.

An image database, which is attached to certain lemmata and objects (i.e. a papyrus or even a tomb), has also been developed, but is still in an experimental stage.[3]

Thesauri with standardized lists of museum collections, dates, provenances, text supports, and grammatical categories supplement these modules. On most levels free text fields are provided for comments.

At present, the text encoding program is installed on the personal computers of the individual collaborators. A finished text or collection of texts will be saved in *XML* format and exported into the joint database *Thesaurus Linguae Aegyptiae* in Berlin.

3. THE CORPUS OF TEXTS

Ideally, the database should contain all the written records of ancient Egypt which are (and become) available. Of course, one generation of scholars cannot feasibly realize such an enormous and time consuming enterprise. For this reason criteria were established for selecting which texts cooperating institutions should encode. A number of corpora were defined according to three basic criteria:

(1) Priority for digitizing would be given to texts which are not represented at all, or are not well represented, in the *Wörterbuch*, usually because they had not been discovered or were not yet published when the *Wörterbuch* itself was published.

(2) Special emphasis would be placed on the Old Kingdom — ideally on all texts from this period, but primarily on tomb inscriptions.

(3) Yet another category comprises well-defined samples of texts from different periods, different regions, and different linguistic phases or with different content.

The most important groups of *Old Kingdom* texts are those in private tombs from the royal cemeteries at *Giza* and *Saqqara* and from the provincial cemeteries at *Akhmim* (el-Hawawish) and *Deir el-Gebrawi*; the Pyramid Texts; rock inscriptions from *Hatnub*, the *Assuan* region, *Nubia*, *Sinai*, and the Eastern Desert; hieratic texts.

To the third category belong letters (Old Kingdom to Third Intermediate Period); selected royal historical and rhetorical inscriptions from the Dynasty XIX; texts from the Amarna Period; texts from the Late and Greco-Roman Periods (ritual texts from private funerary papyri, e.g. the *papyrus of Imuthes* [pMMA 35.9.21], and texts from temple libraries [the Brooklyn Papyri, perhaps from *Elephantine*, and papyri from the temple of *Tebtynis*]).

The first partner to join the project was the *Saxonian Academy of Sciences and Humanities in Leipzig* (*SAW*, since 1999). The *SAW* focuses on hieroglyphic and hieratic literary texts. Since many had already been included in the *Wörterbuch*, the primary task for the Leipzig Research Unit is to make these particularly important texts available in a form which reflects contemporary standards of scholarship. The texts, which range in date from the Middle Kingdom to the Late Period, are categorized as narratives, discourses, wisdom literature, poetry (royal hymns, harpers' songs, love

[3] See Schweitzer, Simon D. 2009. Bildwörterbuch des Ägyptischen: Eine neue Komponente im Thesaurus Linguae Aegyptiae, in: *Göttinger Miszellen* 223, p. 73-79.

poetry, praise of cities), and the so-called Late-Egyptian Miscellanies. In addition, the encoding of Middle Kingdom historical and biographical literature has also begun. The combination of these paraliterary texts with material provided by other partners will contribute to the formation of a balanced corpus of classical Middle Egyptian texts.

In 2000, the scope of the project was significantly enlarged when the *Academy of Sciences and Literature* in Mainz joined with its *Datenbank demotischer Texte* (Database of Demotic texts). As is well known, the compilers of the *Wörterbuch* at first postponed and then later abandoned altogether the intention of integrating Demotic material. Thus it is a great achievement that this artificial gap in the linguistic record of ancient Egyptian is being closed by the database of Demotic texts and the Demotic word list. At the conclusion of the current phase of the project in 2012, the Würzburg-based research unit aims to have made available 80% of all relevant Demotic texts — in other words, more or less all published literary, religious, etc. texts will be included in the database, but only a selection of those repetitive administrative and documentary texts such as contracts, tax receipts, and mummy labels. The texts are categorized by content into literary texts, religious and magical texts, omina and dream texts, administrative and documentary texts, graffiti and inscriptions on objects, scientific literature, school exercises, and *varia*.[4]

In 2003, the *Totenbuch-Projekt Bonn* of the *North Rhine-Westphalian Academy of Sciences, Humanities and the Arts* contributed three complete Eighteenth Dynasty Book of the Dead papyri, one complete papyrus from Dynasty XXI, and another of the Ptolemaic Period. Some additional papyri are represented only by a few spells, to compensate for those missing in the other versions.

The *Digital Heka* project at the *University of Leipzig* (2006-2008) provided a selection of magical and amuletic texts from the Middle and the New Kingdoms. Particularly noteworthy is a collection of magical spells against snakes.

From the *Leuven Online Index of Ptolemaic and Roman Hieroglyphic Texts* (2005-2008) the database received the digitized texts recorded in a number of smaller temples from the Greco-Roman period: *Assuan*, *Bigge*, *Dakka*, *Deir el-Medina* and *Dendur*. The research unit at Leipzig is still editing the data which Leuven provided on the Ptolemaic chapel of *Deir el-Bahri* and on the temple of *Opet* in Karnak.

Wolfgang Schenkel, professor emeritus (University of Tübingen), contributed a preliminary version of his digital corpus of the *Coffin Texts* which has yet, however, to be adapted to *TLA* standards.

3.1. *Texts encoded in the TLA*

The organisation of the *TLA* by contributing institution, rather than according to provenance, type and/or content of the texts, may seem to present a complicated overall picture of the texts which are now available in the database. The following list is intended to provide an impression of the actual content of the *TLA*, but it is neither complete nor exclusive (e.g. demotic texts are not included).

[4] Vittmann, Günter. 2010. Ein neues demotistisches Hilfsmittel. Die ‚Datenbank demotischer Texte‘, in: *Enchoria* 31, p. 144-152.

3.1.1. Compilations of funerary texts

Pyramid Texts	*Unas*, *Pepi I*, *Pepi II*, *Ankhenespepi II*, *Neith* (*BBAW*)
Coffin Texts[5]	Charms against snakes (*Digital Heka*)
Book of the Dead	Eleven 18th Dynasty papyri (*pMaiherperi, pJuja, pNu* complete); six 19th-20th Dynasty papyri; four 21st-22nd Dynasty papyri (*pLondon 10793* complete); two Ptolemaic Period papyri (*pTurin 1791* complete) (*Totenbuch-Projekt*)
Netherworld Books of the New Kingdom	*Destruction of Mankind* (*SAW*)

3.1.2. Literary texts (*SAW*)

The entire corpus of Middle and Late Egyptian literary texts.

3.1.3. Private tomb inscriptions (*BBAW*)

Old Kingdom[6] and First Intermediate Period: *Giza*, *Saqqara*, *Akhmim*, *Deir el-Gebrawi*.

3.1.4. Temple inscriptions from the Greco-Roman period (*Leuven*)

Temple inscriptions from Assuan, Bigge, Dakka, Deir el-Medina, Dendur.

3.1.5. Ritual, religious, and magical texts

Middle Kingdom	Magical texts on papyri and coffins (*Digital Heka*)
New Kingdom	Magical texts on papyri and ostraca (*Digital Heka*)
Greco-Roman Period[7]	Funerary literature (several Osiris or "mortuary" liturgies; *pImouthès* [*pMMA 35.9.21* complete]) (*BBAW*)

3.1.6. Texts from the Amarna Period (*BBAW*)

Tell el-Amarna	Inscriptions on various objects; boundary stelae (A, L, N, S, U)
Thebes	Some parts of the tomb of Kheruef (TT 192); fragments from structures and statues at Karnak; altars (A-E); fragments from the Aten temples; some small objects; stelae and objects of uncertain provenance
Giza	Blocks from the tomb of Ptah-May

[5] As noted above, Schenkel's electronic edition of the *Coffin Texts* is still in an unpublished, preliminary stage.

[6] The royal letters from the tombs Giza 2370, Saqqara LS 16 [S 902] and Qubbet el-Hawa (Assuan) A 8 are found among *Briefe des Alten Reiches und der Ersten Zwischenzeit*.

[7] The embalming ritual (*pBoulaq 3* and *pLouvre 5158*), made available for the *TLA* by *Susanne Töpfer,* is still under revision.

Saqqara	Stela from the tomb of Mery-Neith (Cairo, Egyptian Museum CG 34182)
Heliopolis	Fragment of a statue
Illahun (Gurob)	Base of a statue and three documentary papyri
Hermopolis	Several architectural fragments
Balansoura	Two statues
Gebel el-Silsile	Rock stela of Akhenaten
Assuan	Graffito
Temple of Sesebi	Column with Nefertiti's titles
Sedeinga (Nubia)	Scarab

3.1.7. Historical and biographical texts

Middle Kingdom and Second Intermediate Period	Private biographies of the 11th-12th Dynasty and of the 13th-17th Dynasty (*SAW*)
New Kingdom	Monumental royal texts of the 19th Dynasty from Lower and Upper Egypt (*BBAW*)

3.1.8. Texts from temple libraries (*BBAW*)

Library of an Upper Egyptian temple (Elephantine?), 26th Dynasty-Late Period	Brooklyn Papyri (*pBrooklyn* 47.218.50,[8] 84,[9] 135,[10] 156[11])[12]
Library of the Temple of Soknebtynis at Tebtynis, Roman Period	Daily ritual; mythological manual for the Upper Egyptian nomes; manual for the priest of Sakhmet (papyri from Copenhagen, Florence, Berlin and Oxford)
Library of an Upper Egyptian temple (Abydos?), Ptolemaic Period	Ritual text of *pSalt 825* (*pBM 10090 + 10051*)

3.1.9. Administrative and documentary texts (*BBAW*)

Old Kingdom and First Intermediate Period	Royal letters, private letters and letters to the dead; other documentary texts (mainly pottery from the *Qubbet el-Hawa* near *Assuan*); *Abusir* archive[13]
Middle Kingdom and Second Intermediate Period	Private letters and letters to the dead
New Kingdom and Third Intermediate Period	Royal letters, private letters, letters to the dead and to gods

[8] 'Confirmation du pouvoir royal au nouvel an'.

[9] 'Mythes et légendes du Delta'.

[10] 'A Late Period Hieratic Wisdom Text' (provided by the *SAW*).

[11] 'Le Papyrus magique illustré'.

[12] pBrooklyn 47.218.48+85 ('Un traité égyptien d'ophiologie') is still under revision and will soon be added to the *TLA*.

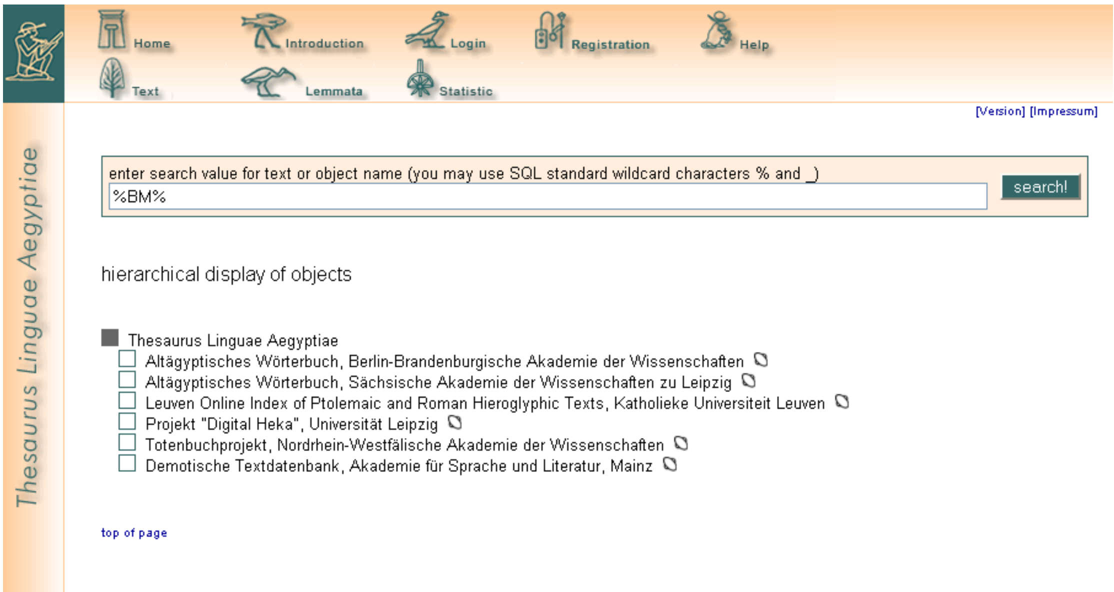
[13] Currently under revision and not yet available.

3.1.10. Graffiti and rock inscriptions (*BBAW*)

Old Kingdom	*Sinai*, *Hatnub*, *Gebel el-Hammam*, *Hagar el-Garb*, *Sehel*; various sites in the wadis of the eastern desert; various sites in *Nubia*

3.2. *Finding texts in the database*

The *TLA Darstellung der Objekthierarchie* (hierarchical display of objects and texts) is organised according to collaborating institution. This provides participating institutions with the means of demonstrating their contribution to the project, when it comes to justifying their funding, since there is no hard copy of the database. It might have been preferable to organise the texts in another way, such as by individual collaborator or present location; by type of text; provenance or date; etc. All of this information is in fact encoded in the database as attributes of each text (*Passportdaten*) and will be fully searchable online in the future. For some types of research it would be useful to be able to set up an individual, purpose-oriented corpus of texts by selecting items from the database. But this is not yet possible.



enter search value for text or object name (you may use SQL standard wildcard characters % and _)

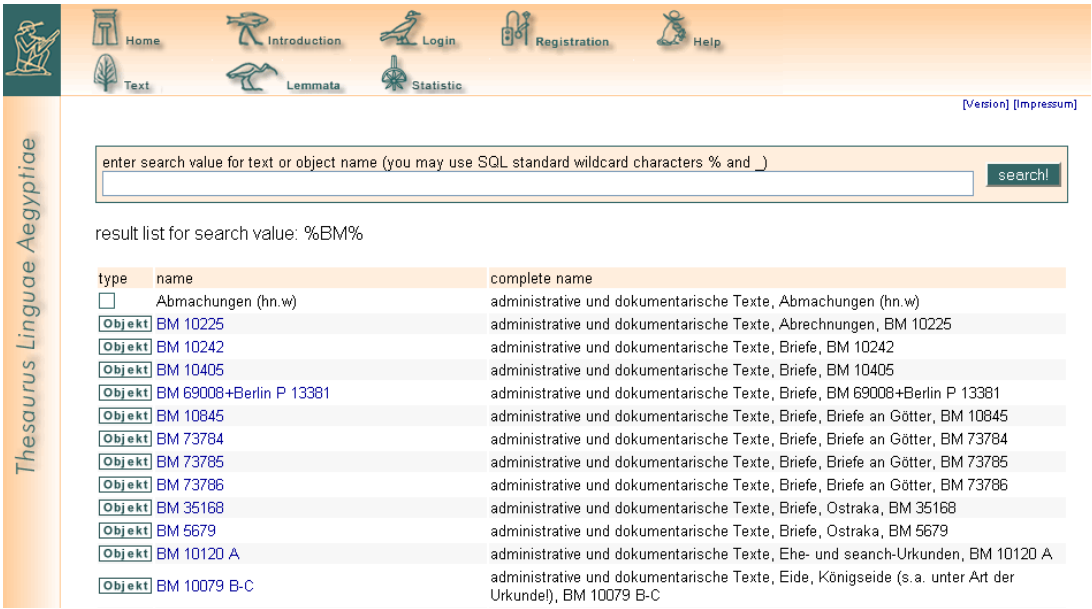
%BM%

hierarchical display of objects

■ Thesaurus Linguae Aegyptiae
- □ Altägyptisches Wörterbuch, Berlin-Brandenburgische Akademie der Wissenschaften
- □ Altägyptisches Wörterbuch, Sächsische Akademie der Wissenschaften zu Leipzig
- □ Leuven Online Index of Ptolemaic and Roman Hieroglyphic Texts, Katholieke Universiteit Leuven
- □ Projekt "Digital Heka", Universität Leipzig
- □ Totenbuchprojekt, Nordrhein-Westfälische Akademie der Wissenschaften
- □ Demotische Textdatenbank, Akademie für Sprache und Literatur, Mainz

Figure 1. Hierarchical display of objects and texts

How can users determine if a given text has been encoded or find it, if they do not know the name of the encoding research unit? For now, it is only possible to search for a text by its name (or part of its name). Each name consists of at least two elements: the designation of the medium on which the text is written and a title or the generally accepted name of the text itself. Since a written record usually belongs to a larger unit, an archaeological complex, or a collection of similar texts, the complete name of a text consists in most cases of a cluster of hierarchical elements which can be searched, even if they are not always visible to the user. These hierarchical elements are labelled *caption*, *group*, *arrangement*, *object*, *part of object*, *scene*, and *text*. The name of the text can contain elements such as the medium, the provenance, owner, present location, inventory number, or a modern name or title of the object or text. When searching for a text, it is in general advisable to preface and close the search-word with the SQL-wildcard "%". For example %BM% should result in a list of British Museum objects contained in the database.



enter search value for text or object name (you may use SQL standard wildcard characters % and _)

result list for search value: %BM%

type	name	complete name
□	Abmachungen (hn.w)	administrative und dokumentarische Texte, Abmachungen (hn.w)
Objekt	BM 10225	administrative und dokumentarische Texte, Abrechnungen, BM 10225
Objekt	BM 10242	administrative und dokumentarische Texte, Briefe, BM 10242
Objekt	BM 10405	administrative und dokumentarische Texte, Briefe, BM 10405
Objekt	BM 69008+Berlin P 13381	administrative und dokumentarische Texte, Briefe, BM 69008+Berlin P 13381
Objekt	BM 10845	administrative und dokumentarische Texte, Briefe, Briefe an Götter, BM 10845
Objekt	BM 73784	administrative und dokumentarische Texte, Briefe, Briefe an Götter, BM 73784
Objekt	BM 73785	administrative und dokumentarische Texte, Briefe, Briefe an Götter, BM 73785
Objekt	BM 73786	administrative und dokumentarische Texte, Briefe, Briefe an Götter, BM 73786
Objekt	BM 35168	administrative und dokumentarische Texte, Briefe, Ostraka, BM 35168
Objekt	BM 5679	administrative und dokumentarische Texte, Briefe, Ostraka, BM 5679
Objekt	BM 10120 A	administrative und dokumentarische Texte, Ehe- und seanch-Urkunden, BM 10120 A
Objekt	BM 10079 B-C	administrative und dokumentarische Texte, Eide, Königseide (s.a. unter Art der Urkunde!), BM 10079 B-C

Figure 2. Search-word with SQL-wildcard: "%BM%"

4. SEARCHING, SORTING, COUNTING AND ANALYSING DATA IN THE *TLA*

The *Thesaurus Linguae Aegyptiae* provides online access to the annotated digital corpus of Egyptian texts, as well as to the material utilized for the *Wörterbuch*. To a certain extent, the new material is linked to the old.

In accordance with the basic structure of the database, access is obtainable through the lexical thesaurus and/or through the corpus of texts. For the present, the image database (cf. §5) is accessed through either one or both; it cannot be used as a search tool since it is still in the experimental stage.

4.1. *Searching, sorting, and counting in the lexical thesaurus*

After entering the database via the lexical thesaurus, the following search operations are available:

- searching the entire corpus for an Egyptian word and its references, either hieroglyphic/hieratic or demotic;
- searching for hieroglyphic writings of lexical entries;
- combined searching for two Egyptian words (or a word and a word class) occurring simultaneously in a specific textual sequence (not necessarily in consecutive order), either hieroglyphic/hieratic or demotic.

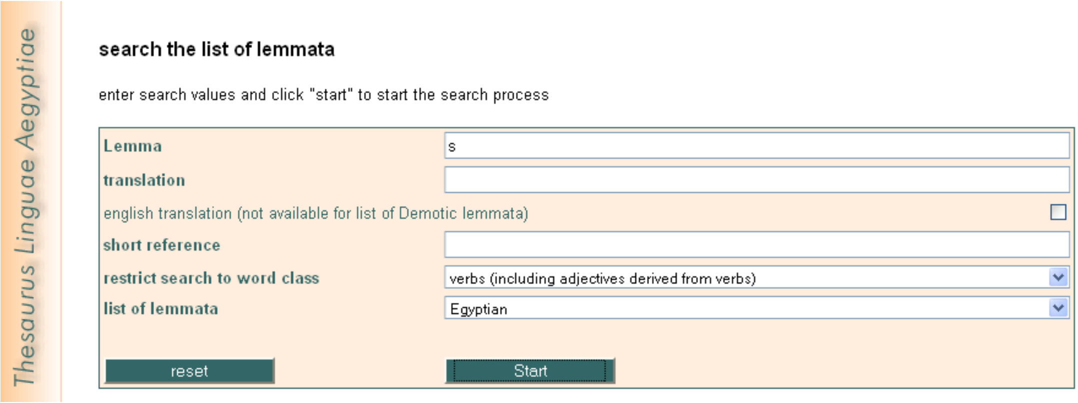
The following statistical analyses can be made:

- collocation analysis for a specific lemma;
- lexical gravity of a specific lemma.

Users can conduct a search for Egyptian words by using the dialogue box "*Lemma*" for a transliteration; the box "translation" for a (German or English) translation; the box "word class" (*Wortart*) by choosing a term from the list (e.g. 'Verbs'); the box "bibliographical short reference" for a bibliographical term (e.g. 'Coffin Texts'); or any combination of these options.

For example, users can search for all verbs (word class) beginning with *s* by writing "s" in the *Lemma* box, or for all personal names that end with *ḥtp.w* by writing "§*-Htp.w$".[14]

[14] Clicking on the Help sign on the page for *searching the list of lemmata* will lead to explanations for using the SQL wildcard characters (§, *, $, etc.) to support search.

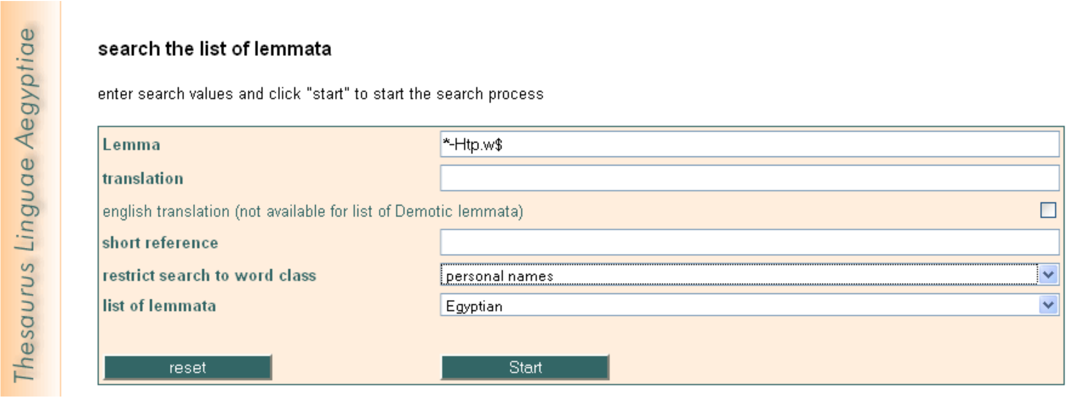


search the list of lemmata

enter search values and click "start" to start the search process

Lemma	s
translation	
english translation (not available for list of Demotic lemmata)	☐
short reference	
restrict search to word class	verbs (including adjectives derived from verbs)
list of lemmata	Egyptian
reset	Start

Figure 3. Search for all verbs beginning with *s*



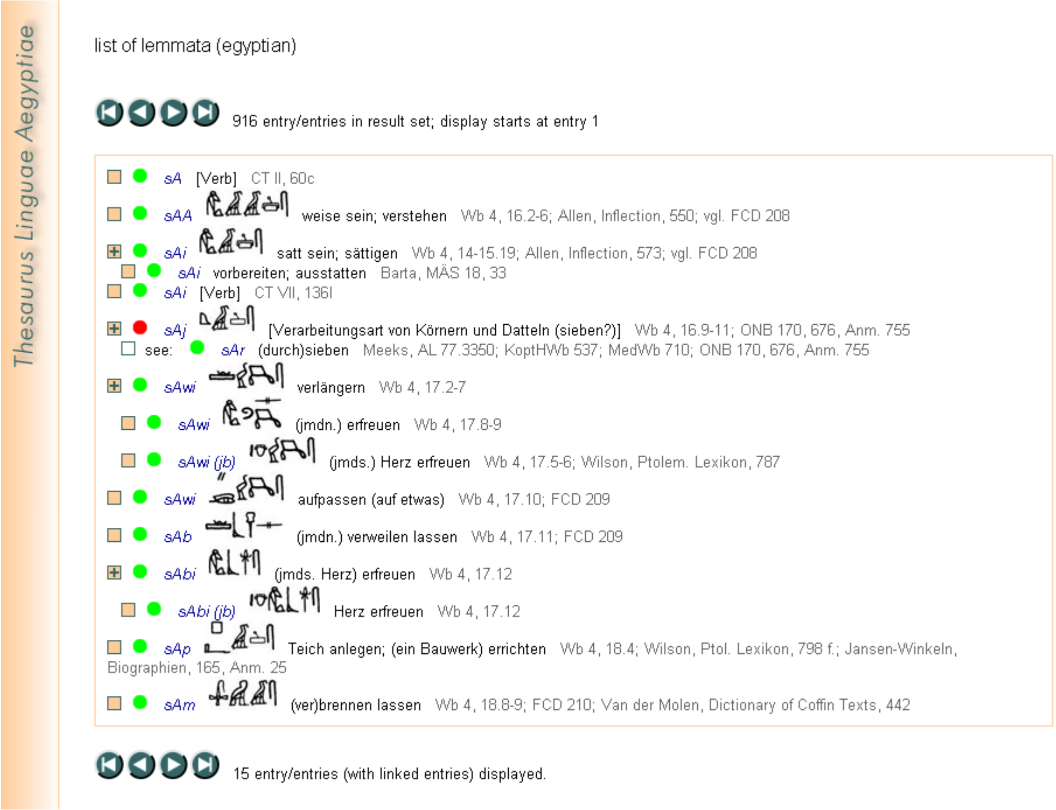
search the list of lemmata

enter search values and click "start" to start the search process

Lemma	*-Htp.w$
translation	
english translation (not available for list of Demotic lemmata)	☐
short reference	
restrict search to word class	personal names
list of lemmata	Egyptian
reset	Start

Figure 4. Search for all personal names that end with *ḥtp.w*

The result will be a list of relevant lemmata, showing the basic hieroglyphic (standard) writing of the entries (if already included in the lexical thesaurus):



list of lemmata (egyptian)

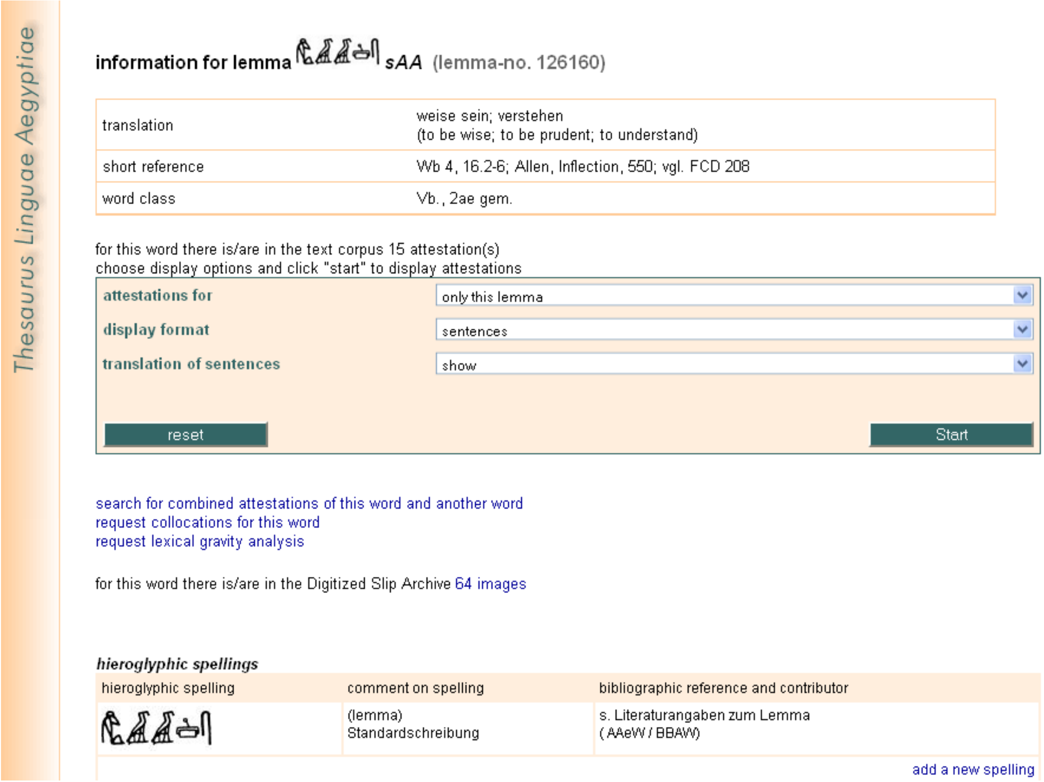
916 entry/entries in result set; display starts at entry 1

- □ ● sA [Verb] CT II, 60c
- □ ● sAA weise sein; verstehen Wb 4, 16.2-6; Allen, Inflection, 550; vgl. FCD 208
- ⊞ ● sAi satt sein; sättigen Wb 4, 14-15.19; Allen, Inflection, 573; vgl. FCD 208
  - □ ● sAi vorbereiten; ausstatten Barta, MÄS 18, 33
- □ ● sAi [Verb] CT VII, 136l
- ⊞ ● sAj [Verarbeitungsart von Körnern und Datteln (sieben?)] Wb 4, 16.9-11; ONB 170, 676, Anm. 755
  - □ see: ● sAr (durch)sieben Meeks, AL 77.3350; KoptHWb 537; MedWb 710; ONB 170, 676, Anm. 755
- ⊞ ● sAwi verlängern Wb 4, 17.2-7
  - □ ● sAwi (jmdn.) erfreuen Wb 4, 17.8-9
  - □ ● sAwi (jb) (jmds.) Herz erfreuen Wb 4, 17.5-6; Wilson, Ptolem. Lexikon, 787
- □ ● sAwi aufpassen (auf etwas) Wb 4, 17.10; FCD 209
- □ ● sAb (jmdn.) verweilen lassen Wb 4, 17.11; FCD 209
- ⊞ ● sAbi (jmds. Herz) erfreuen Wb 4, 17.12
  - □ ● sAbi (jb) Herz erfreuen Wb 4, 17.12
- □ ● sAp Teich anlegen; (ein Bauwerk) errichten Wb 4, 18.4; Wilson, Ptol. Lexikon, 798 f.; Jansen-Winkeln, Biographien, 165, Anm. 25
- □ ● sAm (ver)brennen lassen Wb 4, 18.8-9; FCD 210; Van der Molen, Dictionary of Coffin Texts, 442

15 entry/entries (with linked entries) displayed.

Figure 5. Result for all the verbs beginning with *s*

Clicking on a specific entry will result in a display showing an internal (database) word number or ID, transcription, basic/conventional translation, bibliographical reference, word class, the number of attestations in the digital text corpus, and the number of attestations (images) in the slip archive of the *Wörterbuch*:



information for lemma *sAA* (lemma-no. 126160)

translation	weise sein; verstehen (to be wise; to be prudent; to understand)
short reference	Wb 4, 16.2-6; Allen, Inflection, 550; vgl. FCD 208
word class	Vb., 2ae gem.

for this word there is/are in the text corpus 15 attestation(s)
choose display options and click "start" to display attestations

attestations for: only this lemma
display format: sentences
translation of sentences: show

search for combined attestations of this word and another word
request collocations for this word
request lexical gravity analysis

for this word there is/are in the Digitized Slip Archive 64 images

hieroglyphic spellings

hieroglyphic spelling	comment on spelling	bibliographic reference and contributor
	(lemma) Standardschreibung	s. Literaturangaben zum Lemma (AAeW / BBAW)

Figure 6. Information for a lemma

As transliteration conventions vary, especially as far as endings are concerned, it can be very helpful to enter only the root or the beginning of a word, or to use wildcards.

The attestations already present in the digital text corpus are displayed in chronological order and can be arranged alphabetically, according to the preceding or the succeeding word (*concordance sorted by left or right cotext*). The attestations are rendered contextually, i.e., embedded (and highlighted) in their sentence in the transliteration. The context can be enlarged, and the metadata of the text where the lemma occurs, or the context to which the text belongs (*hierarchical display of objects*), can be called up by clicking on the appropriate button (*Cotext*, *Text*, *Umgebung*). A click on any other word in the transliterated phrase leads to its lemma entry with its hieroglyphic writing, translation, and so forth. Attestations in the digitized slip archive can be limited to the referenced attestations (*Belegstellen*) for the *Wörterbuch der aegyptischen Sprache* or users can choose to have all slips in the archive attesting the relevant lemma included.

It is also possible to search for the co-occurrence of two lemmata in a defined word sequence of a text. The user can define how far they may be separated (up to a maximum of 10 words), regardless of whether one lemma occurs before or after the other, and whether they occur in one sentence or in consecutive sentences. Thus a search for *nḏm* and *jb* would result in both *nḏm jb=f*: "his heart is sweet" and *jb=f nḏm*(*.w*): "his heart has become sweet" (different word order), but also *jw=f nḏm*(*.w*) *ḥr jb=ỉ*: "it is agreeable to my heart" (separated by one word).

A collocation analyser has been implemented in the TLA in order to determine whether some lemmata co-occur more often with a certain lemma than expected. Conducting such a collocation analysis for *jb* and any five words occurring before or after *jb* will provide not only *nḏm*, but also *snḏm*, *swḏꜣ* and *sḥtp* using the statistical measure *T-score*. Alternatively the *MI-score* can be used, but it is less reliable when only a few attestations are involved.[15] To determine the distance to the left or to the right of a given lemma, the "*lexical gravity analysis*" tool may be useful. Taking *snḏm* as an example, collocation analysis shows that *jb* is likely to appear in the first position to the right of *snḏm*.

Especially when using these statistical methods it must be borne in mind that the actual reference corpus (the entire database) is currently not well balanced. On the one hand, many (types of) texts, due to the methods of incorporating data, are still missing, while on the other hand certain texts, such as literary texts and the Book of the Dead, are encoded in more than one version. As a result, the same phrase may occur twice or more times. A tool for merging such manuscripts to create a single — albeit partly artificial — version of the text for statistical purposes has not yet been implemented. Both the missing texts and the multiple versions of some texts bias or falsify statistics based on the entire database. Additional statistical analyses can be performed within the *TLA* on better defined subsections of the database (sections of the text corpus).

4.2. *Searching, sorting, and counting in the corpus of texts*

The organisation of the texts (*hierarchical display of objects*) and the search for text (names) have been dealt with already (§3.2). Once a "family" of texts or a particular text has been selected it is possible to sort and count all the words occurring in it. Behind each group of texts and behind the individual text in the text hierarchy appears the Egyptian hieroglyphic sign 𓐎 for *ḥsb*: "to calculate". A click on this sign opens a window offering:

- three possibilities to create indexes;
- four possibilities to conduct statistical analyses.

Firstly, a word index can be displayed of all words occurring in a group of texts or in a particular text with the number of their attestations within the text(s), organised either alphabetically or by word class. Secondly, an index of names, titles and epithets can be sorted out separately. Thirdly, a frequency index of the words (either all together or organised by word class) including the number of attestations and the corresponding percentage is available under the heading *analysis of the most frequent words*.[16]

The following statistical analyses can also be conducted:

- frequency analysis of the word classes attested in the text(s) providing the number of words within each word class and the frequency of the corresponding words (*word class frequency and type/token statistics*);
- frequency chart of the words can be displayed as a graph and as a numerical spreadsheet (*analysis of word frequency distribution*);
- keyword analysis of all words or of a specific word class can be performed for a given text or corpus in relationship to another corpus (at the moment only for the entire text corpus) (*key words analysis*);
- statistical information is available for individual words within a given (group of) text(s) compared, on the one hand, with the entire corpus and, on the other hand, with the other words belonging to the same word class within the given (group of) text(s) (*statistics for one word*).

[15] Note that the MI-score is the default setting. For most users the T-score will yield results that are easier to interpret.

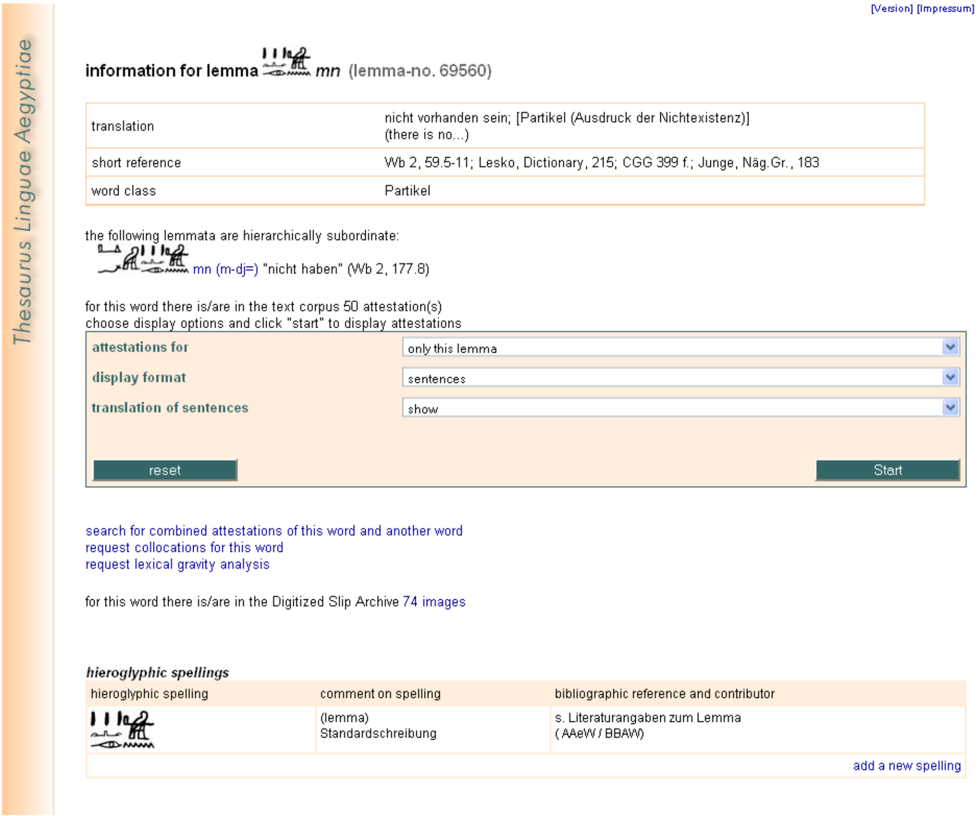
[16] Enter a high number next to the box *maximum number of words in result* if you want to get all words and not just the most frequent ones.

It is recommended, especially for those who are not familiar with statistical methods and terminology, to consult the explanations under the Hilfe button.[17]

5. HIEROGLYPHIC WRITINGS, PICTORIAL DICTIONARY, IMAGE DATABASE

Two further functions are attached to the lexical thesaurus:

- It is possible to search for and add *hieroglyphic spellings* of an individual lemma. The search is accessed through the search field of the lexical thesaurus, but the search for hieroglyphic graphemes cannot be combined with the search by transliteration, translation or word class. When searching for a specific sign or a combination of signs it would be best to read the Help-file first. The use of wildcards will also prove very helpful. If a hieroglyphic writing is missing, it is possible for the user to add a (new) hieroglyphic writing from his home computer into the database.[18] For now, in most cases only a single hieroglyphic writing per lemma is available; for personal names and titles the task of inputting spellings still remains to be done.



information for lemma *mn* (lemma-no. 69560)

translation	nicht vorhanden sein; [Partikel (Ausdruck der Nichtexistenz)] (there is no...)
short reference	Wb 2, 59.5-11; Lesko, Dictionary, 215; CGG 399 f.; Junge, Näg.Gr., 183
word class	Partikel

the following lemmata are hierarchically subordinate:
mn (m-dj=) "nicht haben" (Wb 2, 177.8)

for this word there is/are in the text corpus 50 attestation(s)
choose display options and click "start" to display attestations

attestations for	only this lemma
display format	sentences
translation of sentences	show

search for combined attestations of this word and another word
request collocations for this word
request lexical gravity analysis

for this word there is/are in the Digitized Slip Archive 74 images

hieroglyphic spellings

hieroglyphic spelling	comment on spelling	bibliographic reference and contributor
	(lemma) Standardschreibung	s. Literaturangaben zum Lemma (AAeW / BBAW)

Figure 7. Hieroglyphic spellings in the *TLA*

- In tombs and to a lesser degree on temple walls, captions accompany representations of many objects and activities. A *pictorial or visual dictionary* module is being implemented in *TLA* and will be filled with images for the lexical entries of the lemma list according to the available resources, provided there are no copyright restrictions.[19]

An *image database* attached to the texts is still being tested, using images from the publications of Hermann Junker's excavations at Giza, by kind permission of the copyright holder, the Austrian

[17] The texts under *Help* are currently available only in German but will be translated into English soon.

[18] Unfortunately, a tool for adding hieroglyphic spellings is at present disabled for technical reasons and should *not* be used from outside.

[19] See n. 3.

Academy of Sciences. The image database might later contain plans, photographs, drawings, facsimiles — generally speaking, any pictorial evidence that can help to contextualise and to check the Egyptian text. As some texts are of considerable length it is also planned to attach images to parts of texts, to sentences and even to individual text words. Unfortunately, the copyright restrictions for many images present a considerable obstacle to the further development of this most useful tool.

6. Materials from the Wörterbuch der aegyptischen Sprache

The *Thesaurus Linguae Aegyptiae* also provides access to the digitalised pages of the printed dictionary of Erman and Grapow and to the collection of the 1.2 million alphabetically sorted paper slips which constituted the basic source for the dictionary as published. Additionally, a preliminary manuscript (*Vormanuskript*) of the *Wörterbuch* is available online; compiled between 1906 and 1909, it includes about 800 words. Its interest is not only historical (it shows how at an earlier stage Erman envisaged the scale of the *Wörterbuch*). It also includes culturally relevant information which was lost due to necessary conceptual modifications and to the process of condensation[20] of the *Wörterbuch* as published.

The scanned images of the *Wörterbuch der aegyptischen Sprache* not only allow users to leaf through the volumes "virtually": they also provide access to the digitized collection of paper slips (*Digitalisiertes Zettelarchiv*). Clicking on a *Belegstellennummer* of the *Wörterbuch* leads directly to the corresponding images of the paper slips.

The two other ways to access the digitized slip archive are through the search function of the slip archive itself and through the digital lexical thesaurus of the *TLA*.

7. Short- and long-term perspectives

In the time remaining before termination of the current project in December 2012, work will concentrate on consolidating and implementing those texts and functions already planned including:

- completion of the sub-corpuses of texts that have been encoded into the database;
- implementation of a search function for the metadata (*Passportdaten*);
- implementation of a search mode for the grammatical encodings;
- provision of an English translation for the texts in the Help menu;
- provision of an Arabic version for the user interface.

The objective of a future project would of course be to develop further the functionality of the *Thesaurus Linguae Aegyptiae*. The following list comprises just some possible ways the available data might be expanded, routines already implemented for searching and analysing improved, and new methods and tools developed for examining the diachronic dimension of the 4500 year long history of ancient Egyptian. It would be highly desirable:

- to increase the available data by including additional sub-corpuses of texts which are not yet represented or underrepresented in the text database;
- to integrate the "Egyptian" (hieroglyphic/hieratic) and the Demotic word lists, and to complete the history of the Egyptian language by including and integrating a Coptic word list (and a considerable Coptic text corpus in the text database);
- to improve the visualization of the available analysis routines and implement, e.g., additional methods of text 'mining' or of comparing and clustering texts according to their style, theme or content;

[20] The scale and the degree of detail of the *Vormanuskript* proved to be impracticable for financial reasons.

- to integrate (parts of) the *Multilingual Egyptological Thesaurus* or a similar instrument into the *TLA* and thereby facilitate search operations for text names and metadata in different modern languages;
- to develop new statistical methods and tools for examining the diachronic dimension of ancient Egyptian;
- to establish cooperation and standards for interoperability with other databases of ancient Egyptian such as the Demotic lexicon (*Demotische Wortliste*) of Friedhelm Hofmann. Particularly welcome would be cooperation with the database of *Late Egyptian Texts* at the *University of Liège*, with the *Edfu Temple Project* at the *Göttingen Academy of Sciences and Humanities*, and with the project *Der Tempel als Kanon der religiösen Literatur Ägyptens* at the *Heidelberg Academy of Sciences and Humanities*. Linking the texts in the *TLA* to the digital library *Trismegistos*[21] would also be extremely beneficial.

[21] An interdisciplinary portal of papyrological and epigraphical resources dealing with Egypt and the Nile valley between roughly 800 BC and 800 AD (http://www.trismegistos.org/).

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.

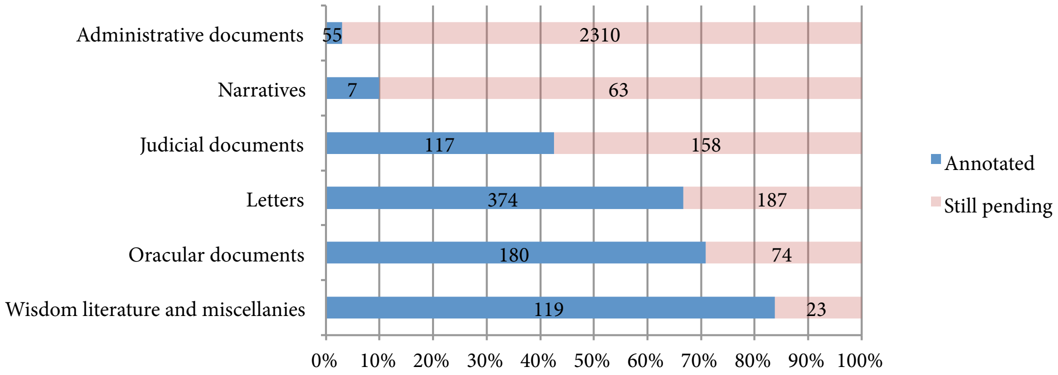


Figure 1. Hieratic documents annotated in Ramses

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.

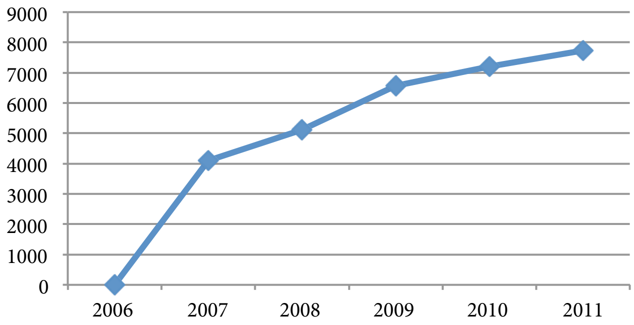


Figure 2a. Lemmata

[11] For these abbreviations, see the bibliography.

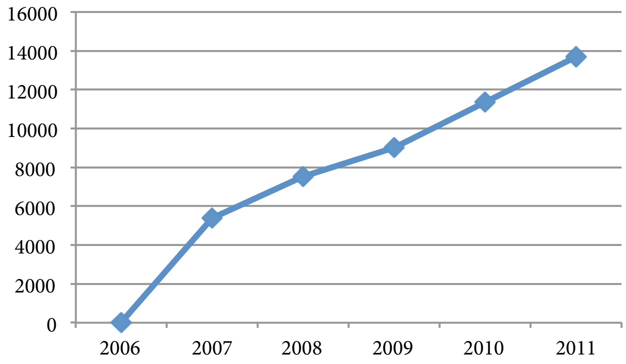


Figure 2b. Inflections

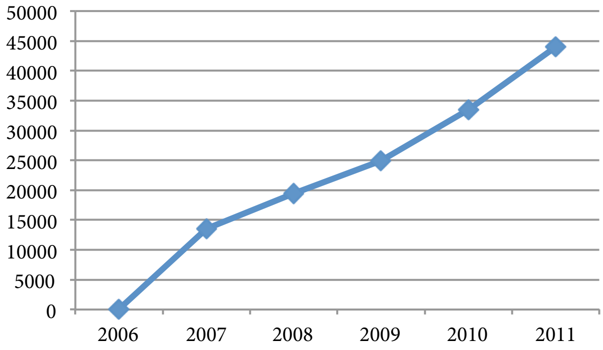


Figure 2c. Spellings

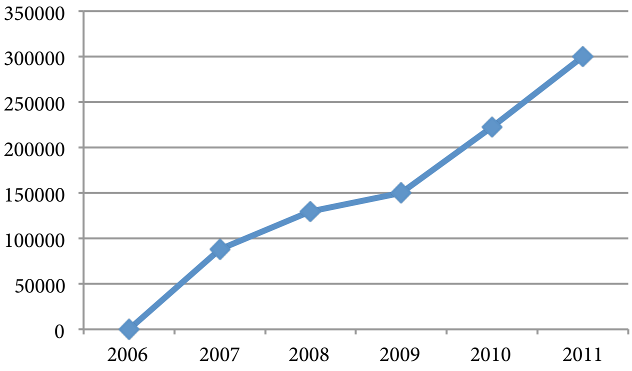


Figure 2d. Words

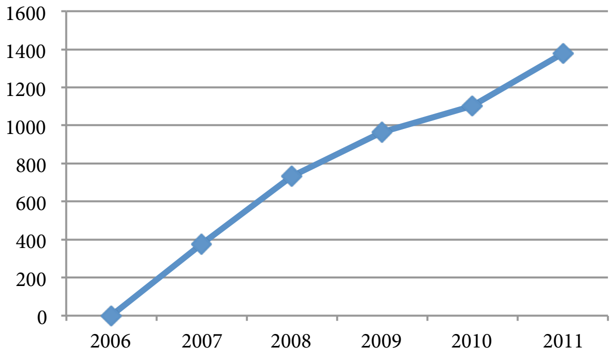


Figure 2e. Texts

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.

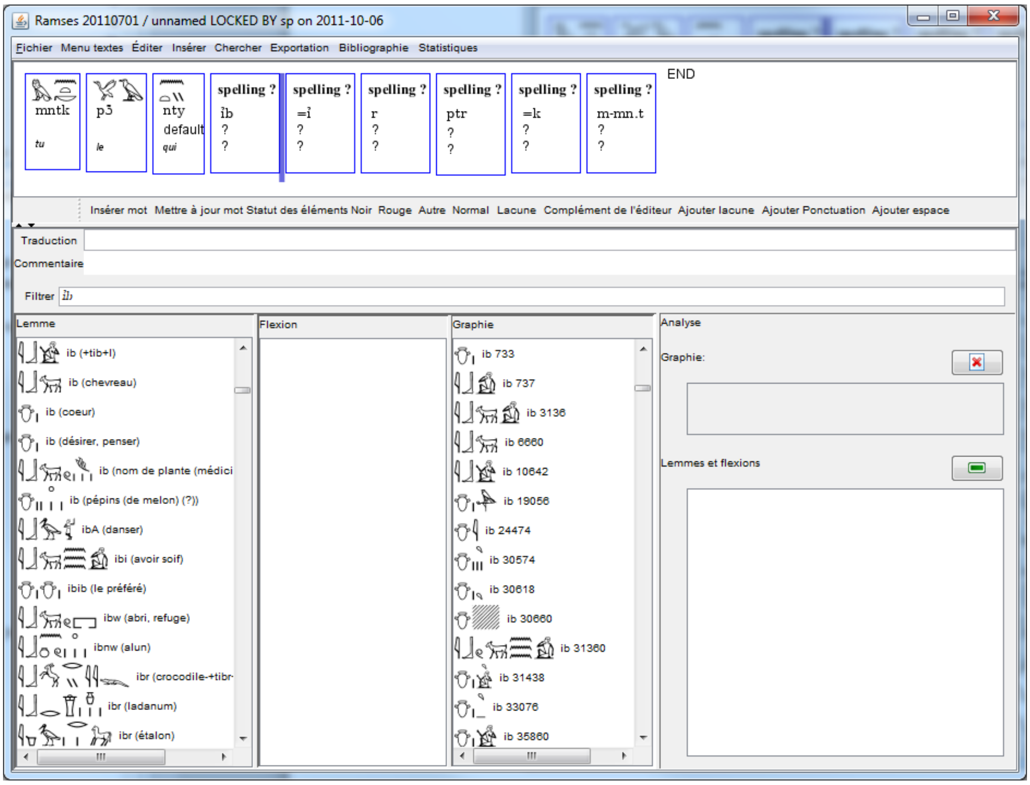
2.3. *The Ramses software*

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).


Ramses 20110701 / unnamed LOCKED BY sp on 2011-10-06
Fichier Menu textes Éditer Insérer Chercher Exportation Bibliographie Statistiques
END
Insérer mot Mettre à jour mot Statut des éléments Noir Rouge Autre Normal Lacune Complément de l'éditeur Ajouter lacune Ajouter Ponctuation Ajouter espace
Traduction
Commentaire
Lemme
Flexion
Graphie
Analyse
Graphie:
Lemmes et flexions

mntk tu | p3 le | nty default qui | spelling ? ỉb ? ? | spelling ? =ỉ ? ? | spelling ? r ? ? | spelling ? ptr ? ? | spelling ? =k ? ? | spelling ? m-mn.t ? ?

Filtrer ỉb

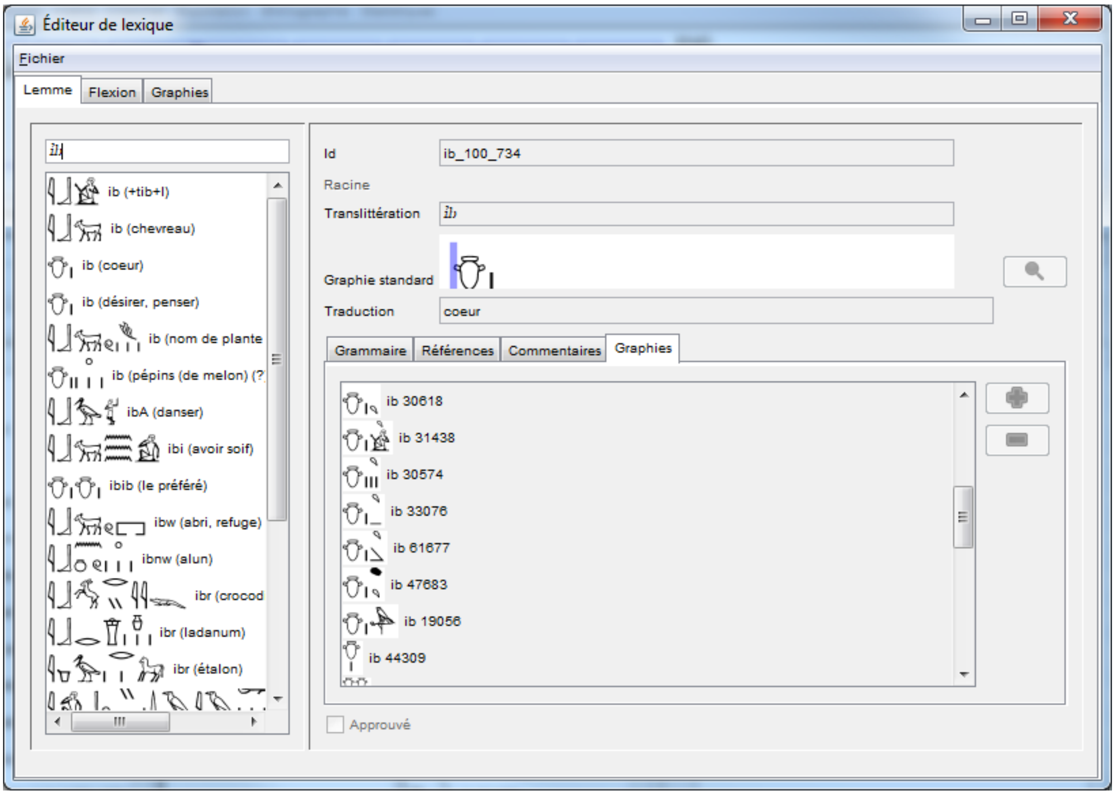
Lemme:
ib (+tib+l)
ib (chevreau)
ib (coeur)
ib (désirer, penser)
ib (nom de plante (médici
ib (pépins (de melon) (?))
ibA (danser)
ibi (avoir soif)
ibib (le préféré)
ibw (abri, refuge)
ibnw (alun)
ibr (crocodile-+tibr-
ibr (ladanum)
ibr (étalon)

Graphie:
ib 733
ib 737
ib 3136
ib 6660
ib 10642
ib 19056
ib 24474
ib 30574
ib 30618
ib 30660
ib 31360
ib 31438
ib 33076
ib 35860

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.



i͗b

ib (+tib+l)
ib (chevreau)
ib (coeur)
ib (désirer, penser)
ib (nom de plante
ib (pépins (de melon) (?
ibA (danser)
ibi (avoir soif)
ibib (le préféré)
ibw (abri, refuge)
ibnw (alun)
ibr (crocod
ibr (ladanum)
ibr (étalon)

Id ib_100_734
Racine
Translittération i͗b
Graphie standard
Traduction coeur

ib 30618
ib 31438
ib 30574
ib 33076
ib 61677
ib 47683
ib 19056
ib 44309

Figure 4. LexiconEditor: Some of the spellings attested for *i͗b* “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.

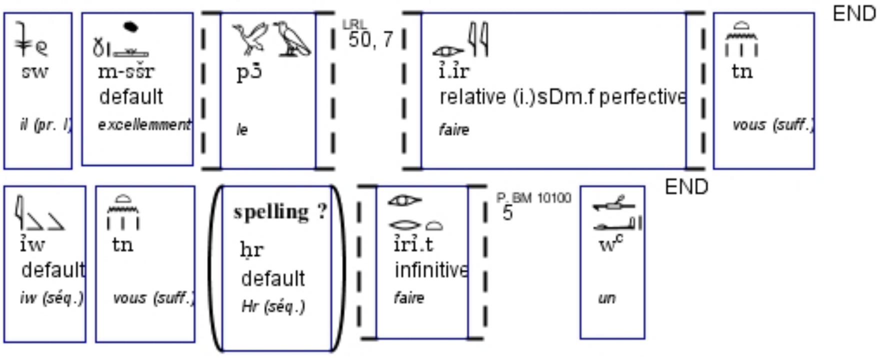


Figure 5. Visualization of tags relating to textual criticism in the TextEditor

(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]

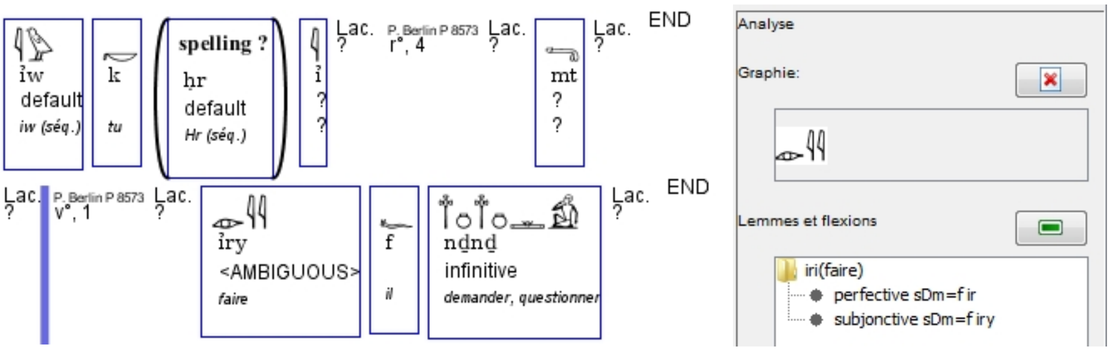


Figure 6. Underspecified annotations and coding of ambiguity in the TextEditor

[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.

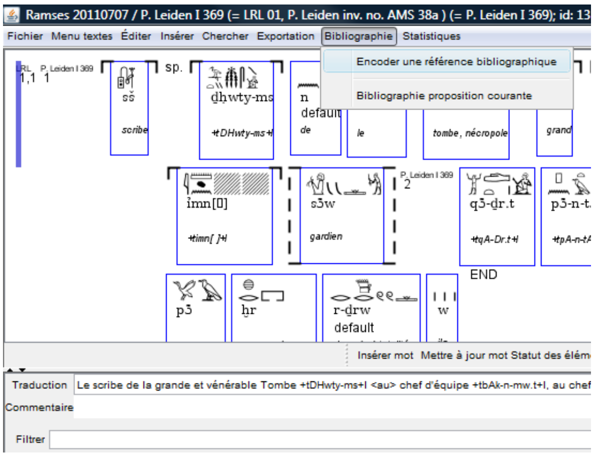
3. TWO NEW FUNCTIONALITIES: RAMSESBIB AND RAMSESEXPORT

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).



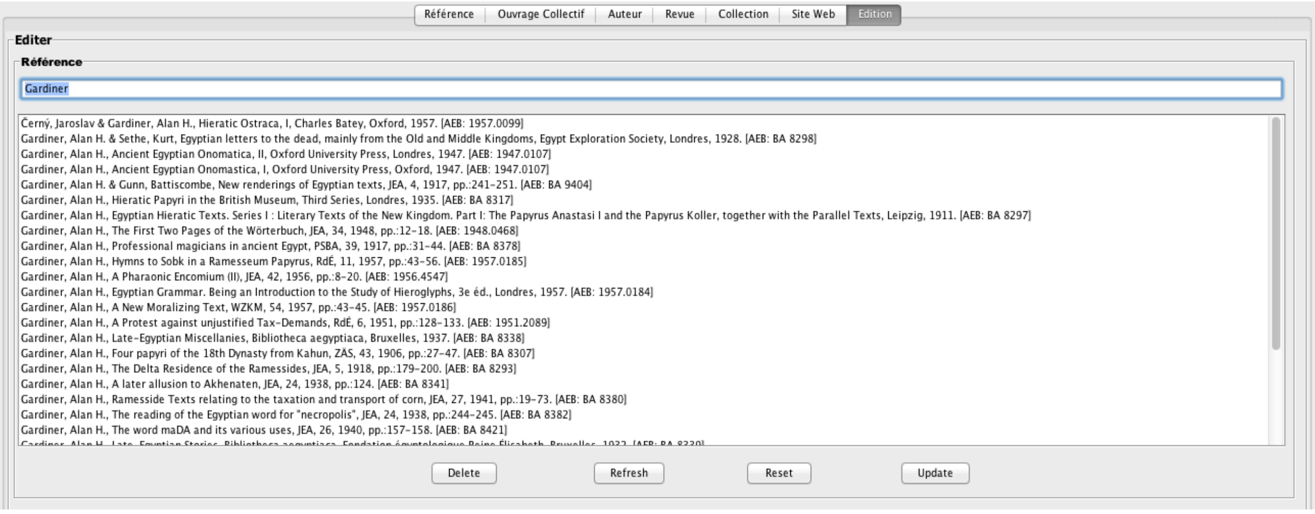
Traduction Le scribe de la grande et vénérable Tombe +tDHwty-ms+l <au> chef d'équipe +tbAk-n-mw.t+l, au chef

Commentaire

Filtrer

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.



Editer

Référence

Gardiner

Černý, Jaroslav & Gardiner, Alan H., Hieratic Ostraca, I, Charles Batey, Oxford, 1957. [AEB: 1957.0099]
Gardiner, Alan H. & Sethe, Kurt, Egyptian letters to the dead, mainly from the Old and Middle Kingdoms, Egypt Exploration Society, Londres, 1928. [AEB: BA 8298]
Gardiner, Alan H., Ancient Egyptian Onomatica, II, Oxford University Press, Londres, 1947. [AEB: 1947.0107]
Gardiner, Alan H., Ancient Egyptian Onomastica, I, Oxford University Press, Oxford, 1947. [AEB: 1947.0107]
Gardiner, Alan H. & Gunn, Battiscombe, New renderings of Egyptian texts, JEA, 4, 1917, pp.:241-251. [AEB: BA 9404]
Gardiner, Alan H., Hieratic Papyri in the British Museum, Third Series, Londres, 1935. [AEB: BA 8317]
Gardiner, Alan H., Egyptian Hieratic Texts. Series I : Literary Texts of the New Kingdom. Part I: The Papyrus Anastasi I and the Papyrus Koller, together with the Parallel Texts, Leipzig, 1911. [AEB: BA 8297]
Gardiner, Alan H., The First Two Pages of the Wörterbuch, JEA, 34, 1948, pp.:12-18. [AEB: 1948.0468]
Gardiner, Alan H., Professional magicians in ancient Egypt, PSBA, 39, 1917, pp.:31-44. [AEB: BA 8378]
Gardiner, Alan H., Hymns to Sobk in a Ramesseum Papyrus, RdÉ, 11, 1957, pp.:43-56. [AEB: 1957.0185]
Gardiner, Alan H., A Pharaonic Encomium (II), JEA, 42, 1956, pp.:8-20. [AEB: 1956.4547]
Gardiner, Alan H., Egyptian Grammar. Being an Introduction to the Study of Hieroglyphs, 3e éd., Londres, 1957. [AEB: 1957.0184]
Gardiner, Alan H., A New Moralizing Text, WZKM, 54, 1957, pp.:43-45. [AEB: 1957.0186]
Gardiner, Alan H., A Protest against unjustified Tax-Demands, RdÉ, 6, 1951, pp.:128-133. [AEB: 1951.2089]
Gardiner, Alan H., Late-Egyptian Miscellanies, Bibliotheca aegyptiaca, Bruxelles, 1937. [AEB: BA 8338]
Gardiner, Alan H., Four papyri of the 18th Dynasty from Kahun, ZÄS, 43, 1906, pp.:27-47. [AEB: BA 8307]
Gardiner, Alan H., The Delta Residence of the Ramessides, JEA, 5, 1918, pp.:179-200. [AEB: BA 8293]
Gardiner, Alan H., A later allusion to Akhenaten, JEA, 24, 1938, pp.:124. [AEB: BA 8341]
Gardiner, Alan H., Ramesside Texts relating to the taxation and transport of corn, JEA, 27, 1941, pp.:19-73. [AEB: BA 8380]
Gardiner, Alan H., The reading of the Egyptian word for "necropolis", JEA, 24, 1938, pp.:244-245. [AEB: BA 8382]
Gardiner, Alan H., The word maDA and its various uses, JEA, 26, 1940, pp.:157-158. [AEB: BA 8421]
Gardiner, Alan H., Late-Egyptian Stories, Bibliotheca aegyptiaca, Fondation égyptologique Reine Élisabeth, Bruxelles, 1932. [AEB: BA 8330]

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.

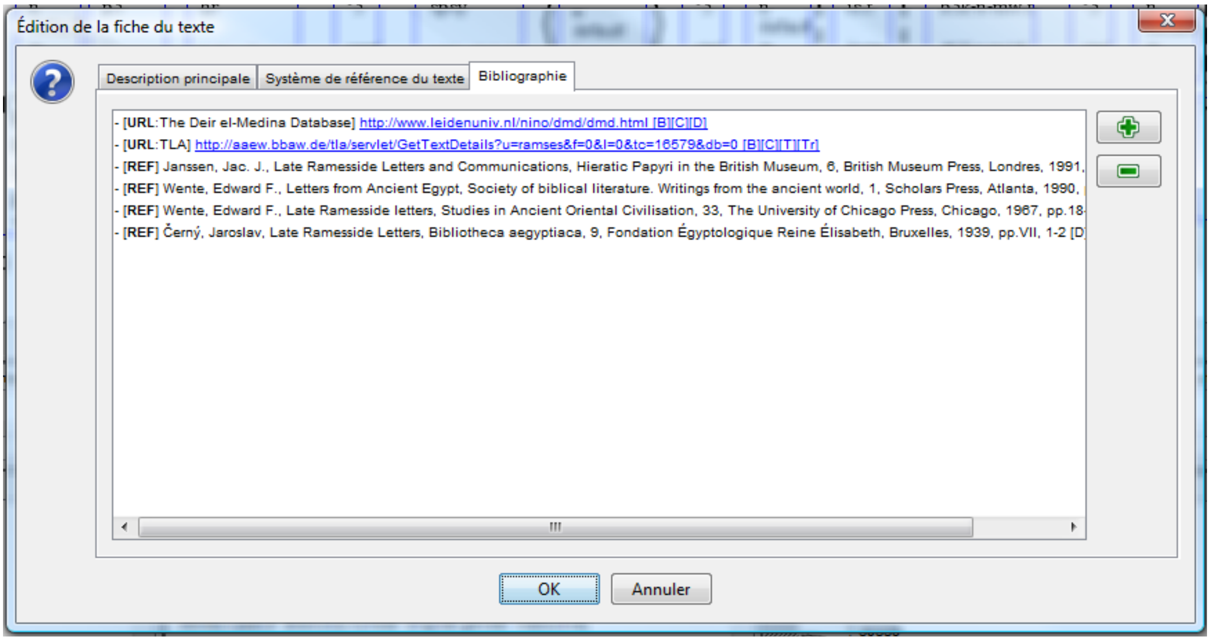
Within the TextEditor, references can be linked to:

(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).

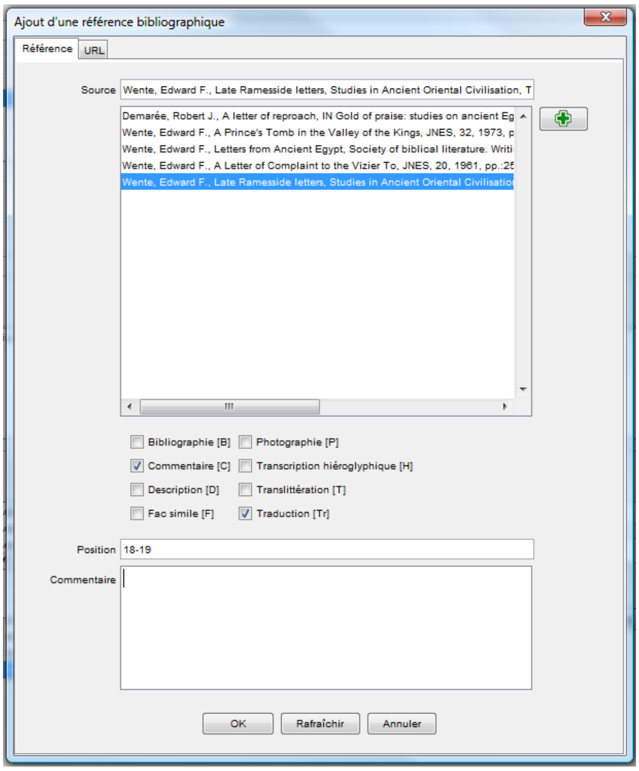
[19] For example the Deir el-Medina Database (http://www.leidenuniv.nl/nino/dmd/dmd.html).



- [URL:The Deir el-Medina Database] http://www.leidenuniv.nl/nino/dmd/dmd.html [B][C][D]
- [URL:TLA] http://aaew.bbaw.de/tla/servlet/GetTextDetails?u=ramses&f=0&l=0&tc=16579&db=0 [B][C][T][Tr]
- [REF] Janssen, Jac. J., Late Ramesside Letters and Communications, Hieratic Papyri in the British Museum, 6, British Museum Press, Londres, 1991,
- [REF] Wente, Edward F., Letters from Ancient Egypt, Society of biblical literature. Writings from the ancient world, 1, Scholars Press, Atlanta, 1990,
- [REF] Wente, Edward F., Late Ramesside letters, Studies in Ancient Oriental Civilisation, 33, The University of Chicago Press, Chicago, 1967, pp.18-
- [REF] Černý, Jaroslav, Late Ramesside Letters, Bibliotheca aegyptiaca, 9, Fondation Égyptologique Reine Élisabeth, Bruxelles, 1939, pp.VII, 1-2 [D]

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.



Source: Wente, Edward F., Late Ramesside letters, Studies in Ancient Oriental Civilisation, T

Demarée, Robert J., A letter of reproach, IN Gold of praise: studies on ancient Eg
Wente, Edward F., A Prince's Tomb in the Valley of the Kings, JNES, 32, 1973, p
Wente, Edward F., Letters from Ancient Egypt, Society of biblical literature. Writi
Wente, Edward F., A Letter of Complaint to the Vizier To, JNES, 20, 1961, pp.:25
Wente, Edward F., Late Ramesside letters, Studies in Ancient Oriental Civilisatio

☐ Bibliographie [B] ☐ Photographie [P]
☑ Commentaire [C] ☐ Transcription hiéroglyphique [H]
☐ Description [D] ☐ Translittération [T]
☐ Fac simile [F] ☑ Traduction [Tr]

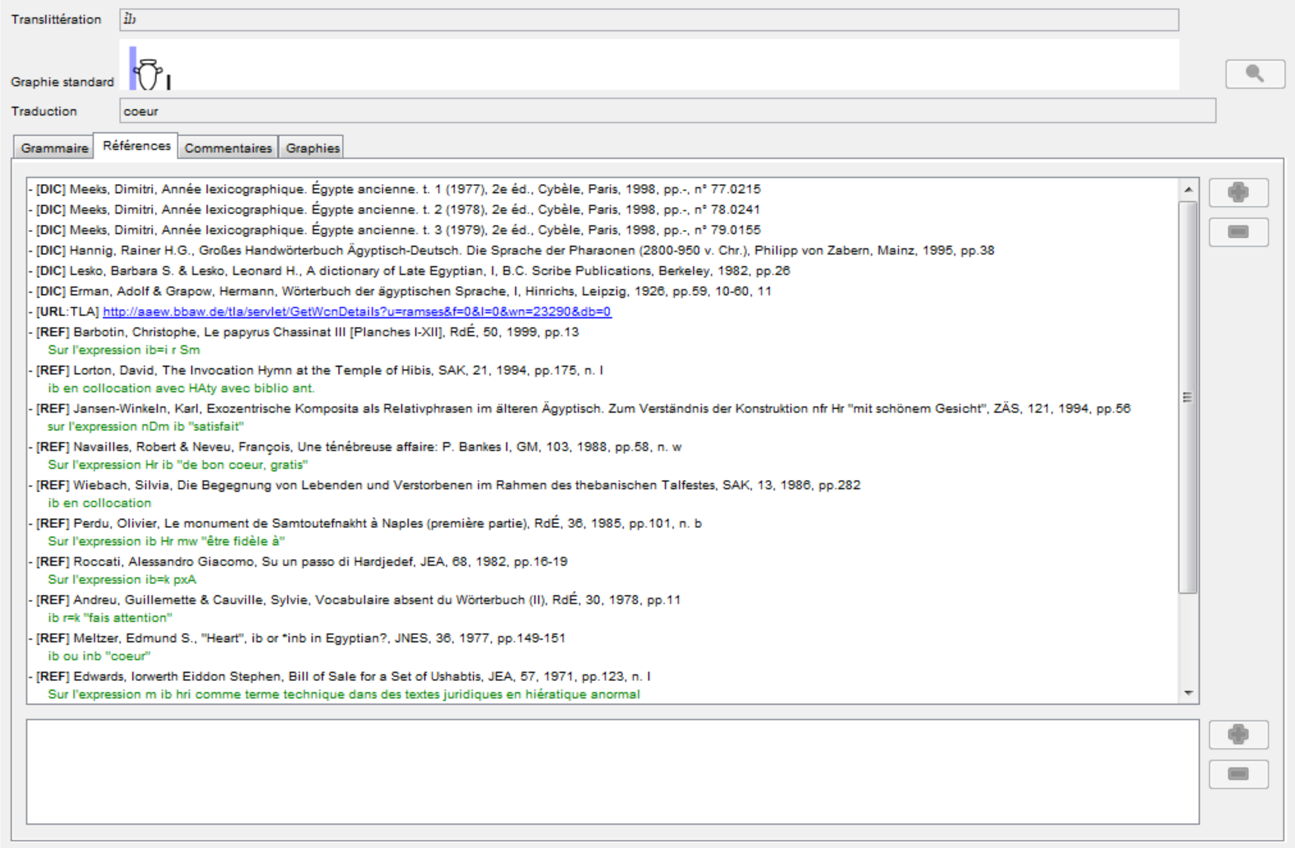
Position: 18-19

Commentaire:

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 *ỉb*). For the sake of readability, references are listed according to these three major groups.



Translittération: ỉb

Traduction: coeur

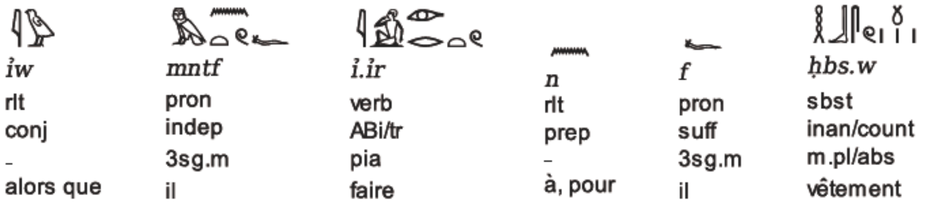
- [DIC] Meeks, Dimitri, Année lexicographique. Égypte ancienne. t. 1 (1977), 2e éd., Cybèle, Paris, 1998, pp.-, n° 77.0215
- [DIC] Meeks, Dimitri, Année lexicographique. Égypte ancienne. t. 2 (1978), 2e éd., Cybèle, Paris, 1998, pp.-, n° 78.0241
- [DIC] Meeks, Dimitri, Année lexicographique. Égypte ancienne. t. 3 (1979), 2e éd., Cybèle, Paris, 1998, pp.-, n° 79.0155
- [DIC] Hannig, Rainer H.G., Großes Handwörterbuch Ägyptisch-Deutsch. Die Sprache der Pharaonen (2800-950 v. Chr.), Philipp von Zabern, Mainz, 1995, pp.38
- [DIC] Lesko, Barbara S. & Lesko, Leonard H., A dictionary of Late Egyptian, I, B.C. Scribe Publications, Berkeley, 1982, pp.26
- [DIC] Erman, Adolf & Grapow, Hermann, Wörterbuch der ägyptischen Sprache, I, Hinrichs, Leipzig, 1926, pp.59, 10-60, 11
- [URL:TLA] http://aaew.bbaw.de/tla/servlet/GetWcnDetails?u=ramses&f=0&l=0&wn=23290&db=0
- [REF] Barbotin, Christophe, Le papyrus Chassinat III [Planches I-XII], RdÉ, 50, 1999, pp.13
  Sur l'expression ib=i r Sm
- [REF] Lorton, David, The Invocation Hymn at the Temple of Hibis, SAK, 21, 1994, pp.175, n. I
  ib en collocation avec HAty avec biblio ant.
- [REF] Jansen-Winkeln, Karl, Exozentrische Komposita als Relativphrasen im älteren Ägyptisch. Zum Verständnis der Konstruktion nfr Hr "mit schönem Gesicht", ZÄS, 121, 1994, pp.56
  sur l'expression nDm ib "satisfait"
- [REF] Navailles, Robert & Neveu, François, Une ténébreuse affaire: P. Bankes I, GM, 103, 1988, pp.58, n. w
  Sur l'expression Hr ib "de bon coeur, gratis"
- [REF] Wiebach, Silvia, Die Begegnung von Lebenden und Verstorbenen im Rahmen des thebanischen Talfestes, SAK, 13, 1986, pp.282
  ib en collocation
- [REF] Perdu, Olivier, Le monument de Samtoutefnakht à Naples (première partie), RdÉ, 36, 1985, pp.101, n. b
  Sur l'expression ib Hr mw "être fidèle à"
- [REF] Roccati, Alessandro Giacomo, Su un passo di Hardjedef, JEA, 68, 1982, pp.16-19
  Sur l'expression ib=k pxA
- [REF] Andreu, Guillemette & Cauville, Sylvie, Vocabulaire absent du Wörterbuch (II), RdÉ, 30, 1978, pp.11
  ib r=k "fais attention"
- [REF] Meltzer, Edmund S., "Heart", ib or *inb in Egyptian?, JNES, 36, 1977, pp.149-151
  ib ou inb "coeur"
- [REF] Edwards, Iorwerth Eiddon Stephen, Bill of Sale for a Set of Ushabtis, JEA, 57, 1971, pp.123, n. I
  Sur l'expression m ib hri comme terme technique dans des textes juridiques en hiératique anormal

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).

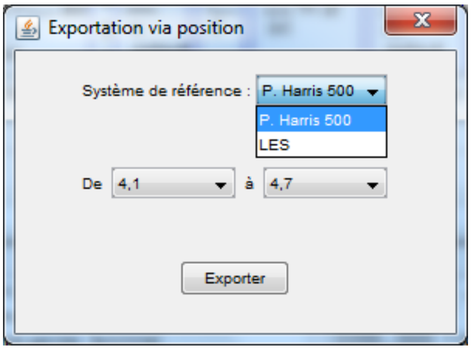


ỉw	*mntf*	*ỉ.ỉr*	*n*	*f*	*ḥbs.w*
rlt	pron	verb	rlt	pron	sbst
conj	indep	ABi/tr	prep	suff	inan/count
-	3sg.m	pia	-	3sg.m	m.pl/abs
alors que	il	faire	à, pour	il	vêtement

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:

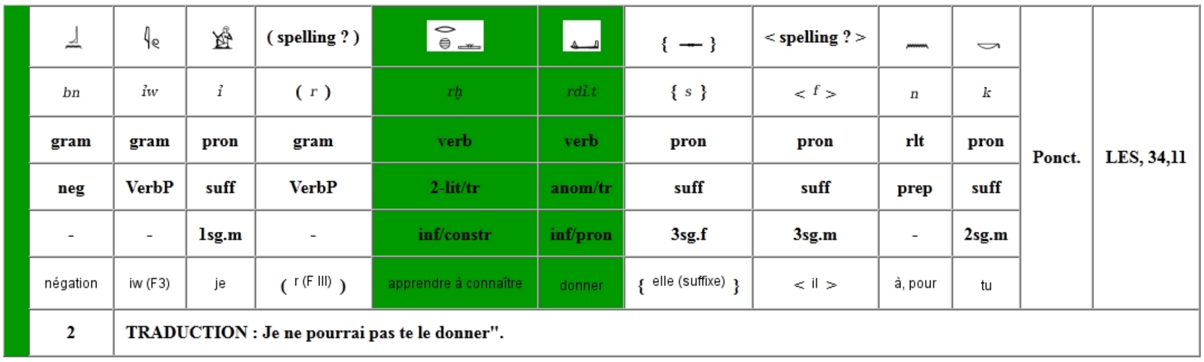


Système de référence : P. Harris 500

De 4,1 à 4,7

Figure 13. Exporting a section of a text using position

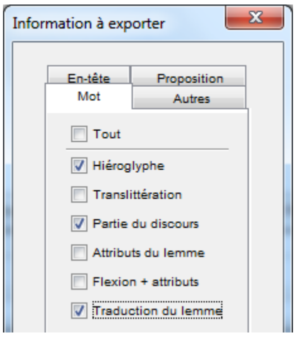
(2) Saving the results of complex searches. With the growing size of the database, some searches already produce a vast amount of results. In order to be able to deal with them properly when studying a given topic, it thus became crucial to enable end-users to export these results in a format convenient 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	*ỉw*	*ỉ*	(*r*)	*rḫ*	*rdỉ.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):



□ Tout
☑ Hiéroglyphe
□ Translittération
☑ Partie du discours
□ Attributs du lemme
□ Flexion + attributs
☑ Traduction du lemme

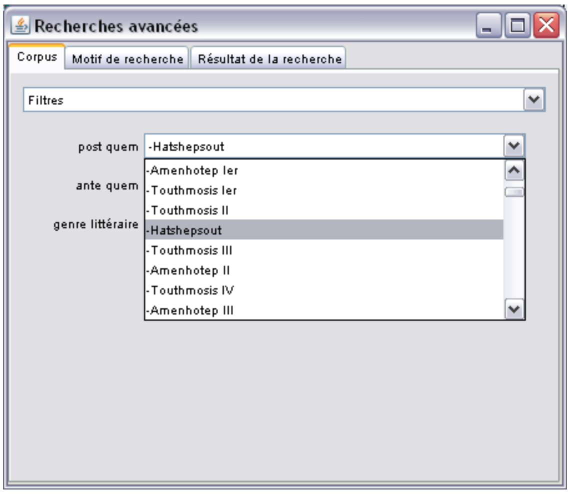
Figure 15. Selection of data to be exported with RamsesExport

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):



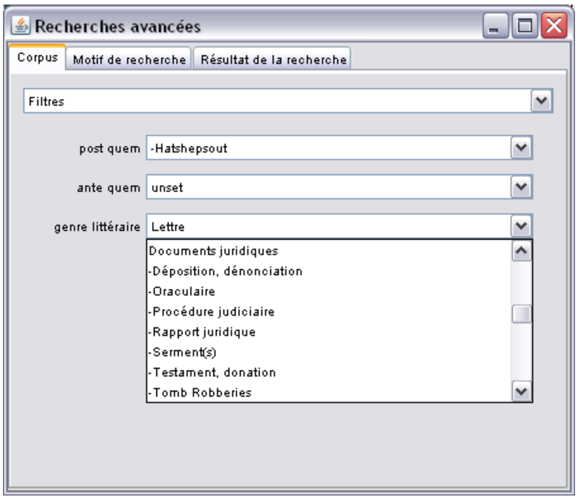
Recherches avancées

post quem: -Hatshepsout

ante quem

genre littéraire

Figure 16a. Selection of a time frame



Recherches avancées

post quem: -Hatshepsout

ante quem: unset

genre littéraire: Lettre

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).

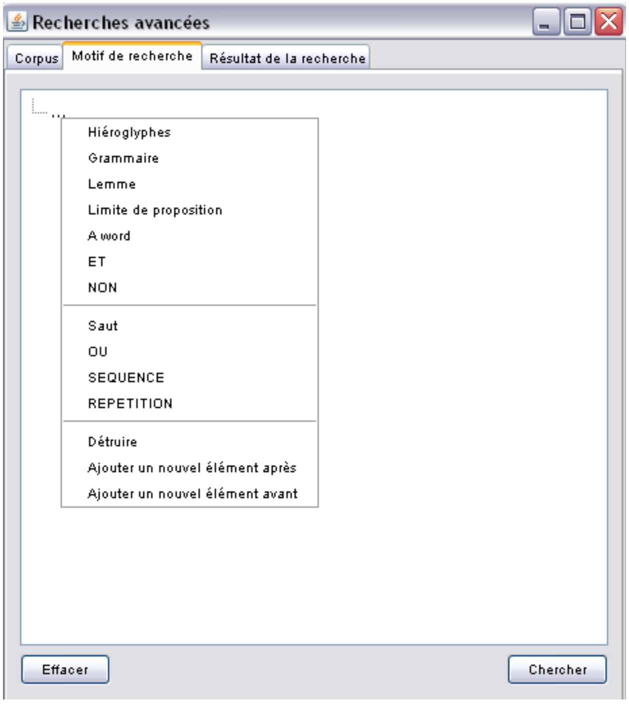


Figure 17. Options of a query

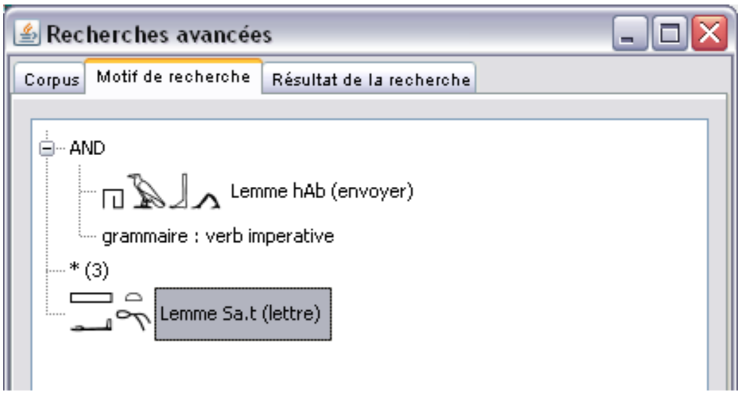
We first give a reminder of the options available for building a query on a single block-occurrence using the different levels of annotation in Ramses. Users can look for (a) a spelling, (b) a lemma or (c) a morphological analysis (and any combination thereof with the operator AND) (Fig. 17); the following 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 *h3b* "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. *h3b* "send" and *šꜥ.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, *h3b* "send" + max. 3 words + *šꜥ.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 *h3b* 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.



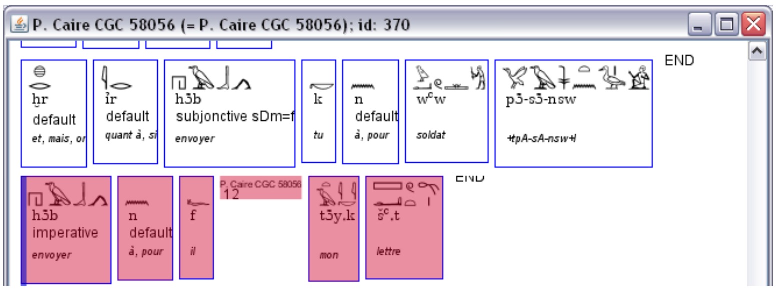
Recherches avancées

Corpus | Motif de recherche | Résultat de la recherche

AND
Lemme hAb (envoyer)
grammaire : verb imperative
* (3)
Lemme Sa.t (lettre)

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

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



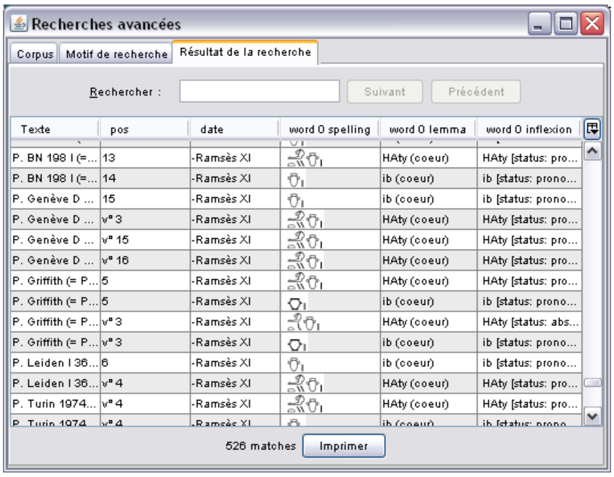
P. Caire CGC 58056 (= P. Caire CGC 58056); id: 370

ḥr default et, mais, or
ỉr default quant à, si
h3b subjonctive sDm=f envoyer
k tu
n default à, pour
wʿw soldat
p3-s3-nsw +tpA-sA-nsw+
END

h3b imperative envoyer
n default à, pour
f il
P. Caire CGC 58056 12
t3y.k mon
šʿ.t lettre
END

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).



Recherches avancées

Corpus | Motif de recherche | Résultat de la recherche

Rechercher :

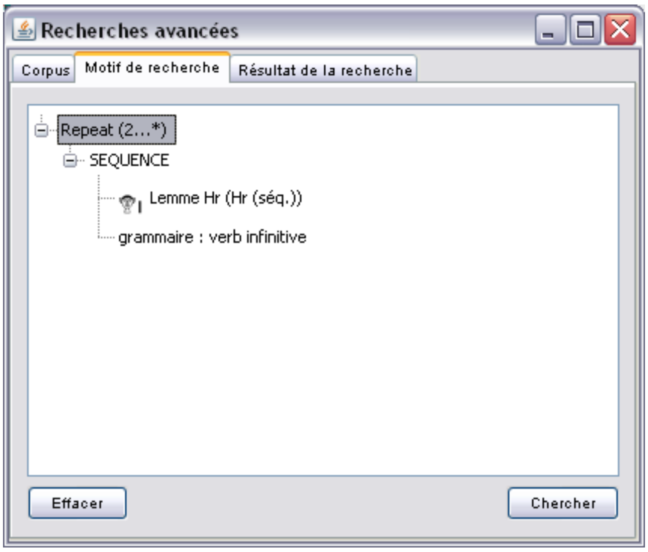
Texte	pos	date	word 0 spelling	word 0 lemma	word 0 inflexion
P. BN 198 I (=...	13	-Ramsès XI		HAty (coeur)	HAty [status: pro...
P. BN 198 I (=...	14	-Ramsès XI		ib (coeur)	ib [status: prono...
P. Genève D ...	15	-Ramsès XI		ib (coeur)	ib [status: prono...
P. Genève D ...	v° 3	-Ramsès XI		HAty (coeur)	HAty [status: pro...
P. Genève D ...	v° 15	-Ramsès XI		HAty (coeur)	HAty [status: pro...
P. Genève D ...	v° 16	-Ramsès XI		HAty (coeur)	HAty [status: pro...
P. Griffith (= P...	5	-Ramsès XI		HAty (coeur)	HAty [status: pro...
P. Griffith (= P...	5	-Ramsès XI		ib (coeur)	ib [status: prono...
P. Griffith (= P...	v° 3	-Ramsès XI		HAty (coeur)	HAty [status: abs...
P. Griffith (= P...	v° 3	-Ramsès XI		ib (coeur)	ib [status: prono...
P. Leiden I 36...	6	-Ramsès XI		ib (coeur)	ib [status: prono...
P. Leiden I 36...	v° 4	-Ramsès XI		HAty (coeur)	HAty [status: pro...
P. Turin 1974...	v° 4	-Ramsès XI		HAty (coeur)	HAty [status: pro...
P. Turin 1974...	v° 4	-Ramsès XI		ib (coeur)	ib [status: prono...

526 matches

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.), *ḥnʿ ntf sḏm* vs. *mtw.f sḏm*, *ỉ.sḏm.t.f* vs. *ỉ.ỉr.t.f sḏm*, etc. This of course also applies for variations at the graphemic level.

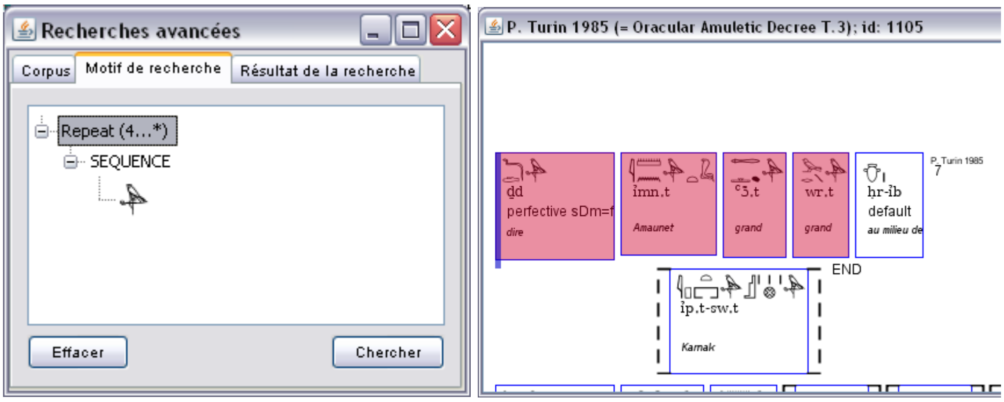
– A new operator that has been added is REPEAT. It enables users to spot contiguous repetitions of lemmata, inflections or graphemes. The number of repetitions can be fixed (with a minimum and a maximum). Consequently, it is possible to test whether there are examples of three adjectives in a row (there 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:



Repeat (2...*)
SEQUENCE
Lemme Hr (Hr (séq.))
grammaire : verb infinitive

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:



Repeat (4...*)
SEQUENCE

P. Turin 1985 (= Oracular Amuletic Decree T.3); id: 1105

ḏd
perfective sDm=f
dire

ỉmn.t
Amaunet

ꜥ3.t
grand

wr.t
grand

ḥr-ỉb
default
au milieu de

P_Turin 1985
7

END

ỉp.t-sw.t
Karnak

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

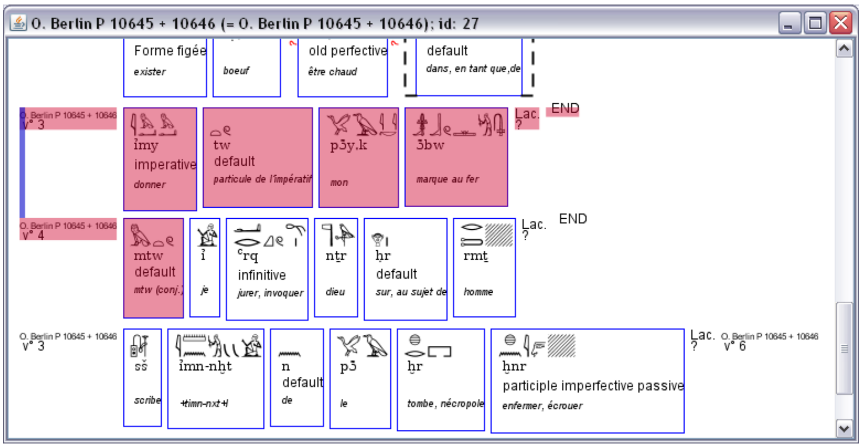
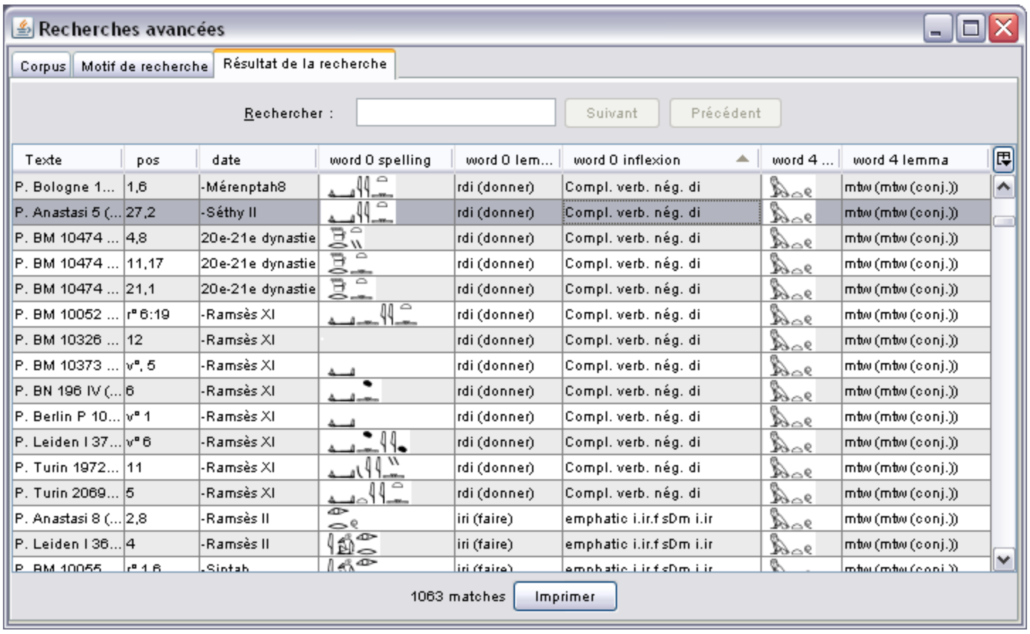


Figure 23. Looking for imperatives followed by conjunctives

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:



Texte	pos	date	word 0 spelling	word 0 lem...	word 0 inflexion	word 4 ...	word 4 lemma
P. Bologne 1...	1,6	-Mérenptah8		rdi (donner)	Compl. verb. nég. di		mtw (mtw (conj.))
P. Anastasi 5 (...	27,2	-Séthy II		rdi (donner)	Compl. verb. nég. di		mtw (mtw (conj.))
P. BM 10474 ...	4,8	20e-21e dynastie		rdi (donner)	Compl. verb. nég. di		mtw (mtw (conj.))
P. BM 10474 ...	11,17	20e-21e dynastie		rdi (donner)	Compl. verb. nég. di		mtw (mtw (conj.))
P. BM 10474 ...	21,1	20e-21e dynastie		rdi (donner)	Compl. verb. nég. di		mtw (mtw (conj.))
P. BM 10052 ...	r° 6:19	-Ramsès XI		rdi (donner)	Compl. verb. nég. di		mtw (mtw (conj.))
P. BM 10326 ...	12	-Ramsès XI		rdi (donner)	Compl. verb. nég. di		mtw (mtw (conj.))
P. BM 10373 ...	v°, 5	-Ramsès XI		rdi (donner)	Compl. verb. nég. di		mtw (mtw (conj.))
P. BN 196 IV (...	6	-Ramsès XI		rdi (donner)	Compl. verb. nég. di		mtw (mtw (conj.))
P. Berlin P 10...	v° 1	-Ramsès XI		rdi (donner)	Compl. verb. nég. di		mtw (mtw (conj.))
P. Leiden I 37...	v° 6	-Ramsès XI		rdi (donner)	Compl. verb. nég. di		mtw (mtw (conj.))
P. Turin 1972...	11	-Ramsès XI		rdi (donner)	Compl. verb. nég. di		mtw (mtw (conj.))
P. Turin 2069...	5	-Ramsès XI		rdi (donner)	Compl. verb. nég. di		mtw (mtw (conj.))
P. Anastasi 8 (...	2,8	-Ramsès II		iri (faire)	emphatic i.ir.f sDm i.ir		mtw (mtw (conj.))
P. Leiden I 36...	4	-Ramsès II		iri (faire)	emphatic i.ir.f sDm i.ir		mtw (mtw (conj.))
P. BM 10055	r° 1,6	-Sintah		iri (faire)	emphatic i.ir.f sDm i.ir		mtw (mtw (conj.))

1063 matches

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.

Building a Construction-Based Treebank of Late Egyptian[*]

The Syntactic Layer in Ramses

Stéphane POLIS & Serge ROSMORDUC

F.R.S.-FNRS (Université de Liège) – Conservatoire National des Arts et Métiers (Paris)

1. INTRODUCTION

The ultimate purpose of the Ramses Project is to provide scholars with a fully annotated corpus of Late Egyptian texts.[1] Unsurprisingly, the annotation of the corpus with syntactic structure came as the last significant development of the project.[2] This part of the software — called SyntaxEditor — has been (and still remains to some extent) an actual challenge in its own right; indeed, several requirements had to be handled simultaneously as regards (1) the syntactic formalism to be implemented and the related representational format, (2) the specificities of the annotation scheme to be developed, and (3) the ergonomic demands of annotation. These needs can be summarized as follows:

(1) From a linguistic viewpoint, the syntactic formalism had to be as theory neutral as possible — i.e. free from theoretical idiosyncrasies, with the evident goal of ultimately allowing scholars from diverse backgrounds to retrieve data on Late Egyptian syntax profitably; at the same time, the generic nature of this formalism could not lead to a simplification of the syntactic annotation: the diversity of the syntactic facts found in the Late Egyptian corpus had to be handled and annotated in its complexity. Besides the traditional specification of "groups" or constructions[3] and accepted part-whole structure hierarchies of constructions, syntactic functions (or roles) — and, crucially, not abstract syntactic relations — have to be explicitly defined for any element according to the construction it belongs to.[4] Furthermore, the corpus has to be annotated not only for skeletal syntactic structure (the so-called "bracketing" task): we wanted a representational model that handles discontinuous constituents on the one hand and that allows, on the other hand, for the annotation of "horizontal" relations between

[*] Serge Rosmorduc is responsible for all IT conception and development to date. The theoretical principles that lie behind the implementation of the SyntaxEditor have much benefited from the expertise of Eitan Grossman who held a post-doctoral position within the Ramses project in 2009-2010. Our thanks are due to Todd Gillen for proofreading the English.

[1] See Polis, Honnay & Winand in the current volume.

[2] As far as the implementation of the annotating tools is concerned, at least. Indeed, the Web application that will give the community of both linguists and Egyptologists access to the corpus is still to be developed.

[3] In this paper, the label "group" is understood generically as referring to any kind of construction at the lexical, idiomatic, phrasal, clausal and textual levels. As "groups", constructions can be compared to constituents in constituency-based formalisms and to a head-dependent(s) relation in dependency-based formalism.

[4] On this basic principle, see *inter alii* Croft 2001: 5.

constructions or elements of constructions (the domains related to textual cohesion — e.g. co-indexation of pronouns and nouns phrase, co-reference, etc.[5] — and information structure).

(2) The SyntaxEditor software had to be abstract enough to allow for an Egyptological team to perform by itself — i.e. in a self-standing (and intelligible) definition file — different types of changes within an annotation scheme that is likely to develop considerably as new constructions are encountered. The goal is indeed not to write *a priori* a grammar in the annotation scheme, but to facilitate the later writing of a grammar based on the documented constructions in the corpus. At the same time, the annotation scheme had to be developed in a way to constrain somehow the annotating process (in order to ensure the coherence of the encoding) and to control beforehand (and to facilitate thereby) the annotators' work.

(3) As for the ergonomics, unlike in modern language corpora where the bracketing task is usually performed with (deterministic) parsers and where, ideally, the parser's output is hand-corrected by annotators in a second step,[6] no parser is available from scratch for Late Egyptian texts. If such a tool is to be part of long-term plans,[7] the ergonomics of the annotation tools had to be designed in a way that would (a) take advantage of the pre-existing annotations (part-of-speech tagging, lemmatization, morphological analysis and translation) and (b) make the chunking of sentences into constructions and the analysis of functions (or roles) of the grouped elements and constructions as quick and straightforward as possible.

This paper addresses these issues by reviewing the current state of the annotation tool, i.e. the SyntaxEditor. In a first section (§2), we argue in favor of a construction-based syntactic formalism, i.e. a formalism that is neither framed in a constituency- nor in a dependency-based model, and aims at encoding the widest varieties[8] of syntactic constructions without positing (in advance) abstract syntactic functions. In the following section (§3), we introduce the evolutionary annotation scheme: it is written in a Syntax Description Language (SDL) and can easily be modified by the annotators any time a previously unattested construction pops up in a texts (without further programming work involved). In the next section (§4), we succinctly describe the steps for manually annotating a text with syntactic structures and present the capabilities of the SyntaxEditor. Finally (§4), we broach future developments: the search engine as well as an interactive parser sensitive to mark-up data.

2. The syntactic formalisms

At present, the Ramses corpus contains a reasonable number of lemmatized and morphologically annotated texts (as of late 2011, ca. 1400 texts for a total of 300000 words). We may therefore proceed with the last significant step of the first phase of the project as regards the annotation procedure, i.e. providing the corpus with a full syntactic analysis.

While the software developed for encoding the lemmatised texts was designed as a one-dimensional linear system — the texts were analysed word by word, each word being assigned a spelling, a lemma and an inflexion —, the syntactic layer calls for some kind of tree editor: in any kind of approach to syntax, the elements of sentences are at some point hierarchically ordered in a two dimensional graph.

[5] For the domain of coreference information in corpus linguistics, see e.g. the Potsdam Commentary Corpus (Stede 2004; other examples cited in Dipper & Götze 2005) or the Spanish CESS-ECE corpus (Recasens *et al.* 2008).

[6] As for example in the Penn Treebank (see Marcus *et al.* 1993: 313-314).

[7] See §5 for an alternative view.

[8] It is worth noting that the syntactic encoding of syntax in Ramsès is *emic* in the sense that it does not take into account allographies and allomorphies that are dealt with in the Text/LexiconEditors, see Polis, Honnay & Winand in the current volume.

2.1. *Dependency-based grammars and phrase structure grammars*

Basically, we had the choice between two overwhelmingly dominant families of formalism in corpus linguistics: dependency grammars[9] and the phrase structure grammars (or constituency grammars).[10]

In dependency grammars — that go back, in modern times, to the *Éléments de syntaxe structurale* of Lucien Tesnière[11] —, there are no phrasal nodes. Everything is modelled as asymmetrical dependency relationships between words: heads (governors of each structure) and dependents. Accordingly, a sentence such as 𓆓𓂧𓏲𓏥𓅓𓂋𓏤𓌡𓂝𓏤 *ḏd=w m r3 wꜥ* "they said with one mouth" could be represented as:

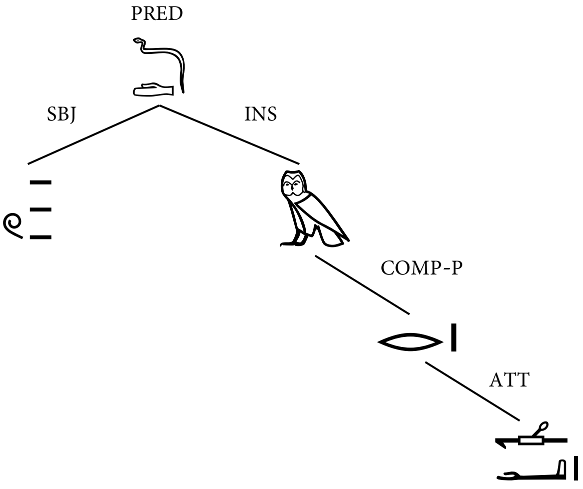


Figure 1. Basic dependency analysis

In our opinion, two obvious advantages of the dependency-based formalism are (1) that, unlike in constituency formalism, syntactic functions are always spelled out explicitly through the specification of the relation types, and (2) that valency patterns or argument structures (especially of verbal predication) are more directly retrievable.

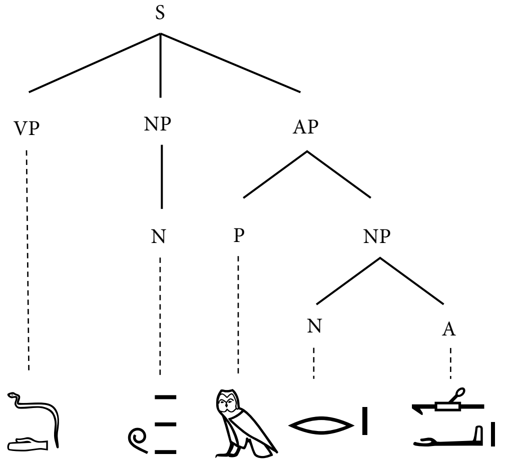


Figure 2. Basic phrase structure analysis

The phrase structure grammar, on the other hand, would group the words in phrases, themselves grouped in higher level phrases, up to the sentence level. Instead of a one-to-one relationship between

[9] An example is the Prague Dependency Treebank for Czech (see Hajic 1999), admittedly a pioneer in this domain.

[10] Much of the work on Treebanks focuses primarily on modern languages. Treebanks of ancient text languages remain rather uncommon; see however McGillivray *et al.* (2009) and Haug *et al.* (2009) as well as references 8-14 cited in Bamman & Crane 2011.

[11] More recently, see inter alii Mel'čuk 1988; Polguère & Mel'čuk 2009.

the words of a sentence and the nodes in the syntactic tree, there is a one-to-one-or-more correspondence between elements of a sentence and syntactic nodes, as exemplified in Fig. 2.

A definite advantage of constituency-based formalisms — at least for an ancient language like Late Egyptian whose texts are often fragmentary — is that it is easy to define a group without describing its entire structure (for instance, to say that *m r3 wꜥ* "with one mouth" in Fig. 2 is an adverbial phrase without analyzing further its constituency). In dependency grammars, the equivalent would be to create temporary unlabelled links between the three words, which is obviously less convenient, especially for complex phrases.

2.2. *A third way: A construction-based Treebank*

Both approaches have obvious pros and cons depending, first, on each scholar's theoretical assumptions regarding syntactic structures, of course, and — more practically — on the language being syntactically annotated.[12]

Given the fact that our basic requirements were (a) to be as theory neutral as possible while taking into account the diversity of the syntactic facts found in the Late Egyptian corpus, (b) to make the annotation of functions explicit[13] in each individual syntactic environment, and (c) to allow for annotations of horizontal relations between elements (graph relations), we tried to combine the advantages of both dependency-based grammars and phrase structure grammars and to overcome what we consider to be their respective shortcomings by developing a simple and intuitive *construction-based* formalism.[14] It should be stressed here that this is much in agreement with the practice in Egyptological linguistics that has traditionally been "Construction Grammar" *avant la lettre*, e.g. with the identification of numerous *patterns*. By doing so, we do not exclude output formats such as dependency or phrase structure graphs, but we envision them as two possible export formats of a more generic formalism[15] that allows describing the syntactic structures in their complexity at the level of surface forms, i.e. not at the level of posited deep structure.[16]

Based on some fundamental tenets of Construction Grammar[17] (CxG) and after a close look at innovative tools such as EMDROS[18] and Notabene,[19] we decided to use the following formalism: the analysis will consist of a set of syntactic constructions, called "groups" in the annotation scheme. A group represents any syntactic construct, from complex words,[20] idioms and simple phrases (like noun or adjectival phrases), to sentences (with various types of predications — including, crucially for a language like Late Egyptian, non-verbal predication patterns — and argument structure schemata) and even paragraphs or entire texts. A group in the formalism (i.e. a construction from a linguistic point of view) has the following properties:

[12] Dependency grammars, for example, prove to be easier to implement for languages involving a relatively free word order.

[13] In this respect, see the remarks in Blache 2000: 85.

[14] It turned out that the annotation scheme of the TIGER Treebank for German (see Brants *et al.* 2002) is actually close to principles advocated for in the present paper.

[15] On the links and transformations between dependency-based and constituency-based graphs, see Robinson 1970 and Mazziotta 2010b: 144, which was our prime source of inspiration in this respect.

[16] Therefore, we do not have to posit (frightening) null elements in the tagset.

[17] See especially Croft 2001; Goldberg 1995, 2006 (with the literature cited p. 18-19). The surface generalizations or the "what you see is what you get" approach to syntactic form that is adopted in CxG is particularly worth mentioning this context, for it does not derive one construction from another and it avoids positing zeroes in the syntactic analysis.

[18] See Petersen 2004.

[19] See Mazziotta 2010b.

[20] The morphemic and word levels, although an obvious part of any analyzable construction is dealt with at the level of the LexiconEditor.

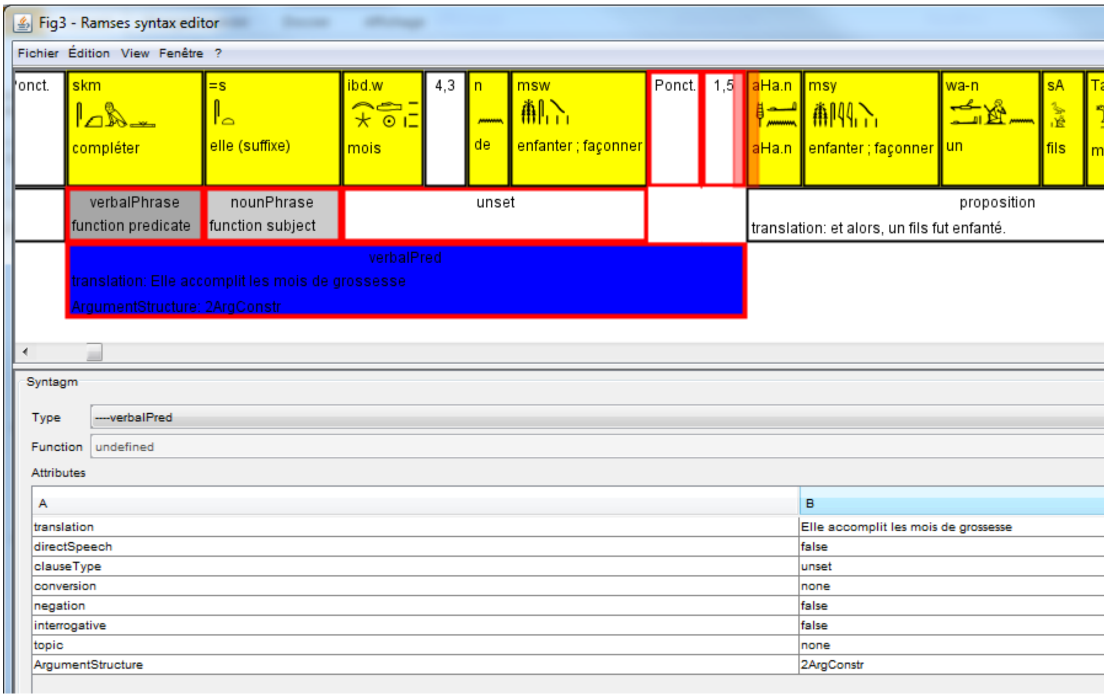
(1) A group may contain one or several basic elements and/or other groups. The functions of children elements and groups depend on the type of parent construction. We see, for example, no point in assuming categories like 'subject' across clause types: such categories must be the object of study based on the annotated data. Therefore, each construction type has its own features and syntactic function. For instance, the S function or role, i.e. 'intransitive subject' is only possible for an element or construction that is part of a higher level construction identified as an intransitive construction.

(2) A group can have various attributes (meant to capture and annotate the combination of different constructions in a single group For instance, an example such as *in wn di=f is.t ḫ3rw* "does he have a crew of Syrians" [*LES* 67,3-4] is an existential-possessive construction that has the attribute 'question construction').

(3) A group need not be continuous; it can be discontinuous.

Moreover, not all links between groups are of a hierarchical parent/child nature: other "horizontal" links are possible in the formalism (graph and not tree type) in order to represent inter alia the phenomena related to textual cohesion and information structure, such as anaphoric relations.

This formalism is admittedly very loose. The specification of the function of a phrasal construction in a sentence, for instance, is not mandatory. That way, partial analysis can be built, which — as already stated above — was a basic requirement for a text language in which many documents are fragmentary. Ideally, we would subsequently run some checking software on the analysis so as to detect what has been left under-specified.



translation: et alors, un fils fut enfanté.

verbalPred
translation: Elle accomplit les mois de grossesse
ArgumentStructure: 2ArgConstr

Syntagm

Type ----verbalPred

Function undefined

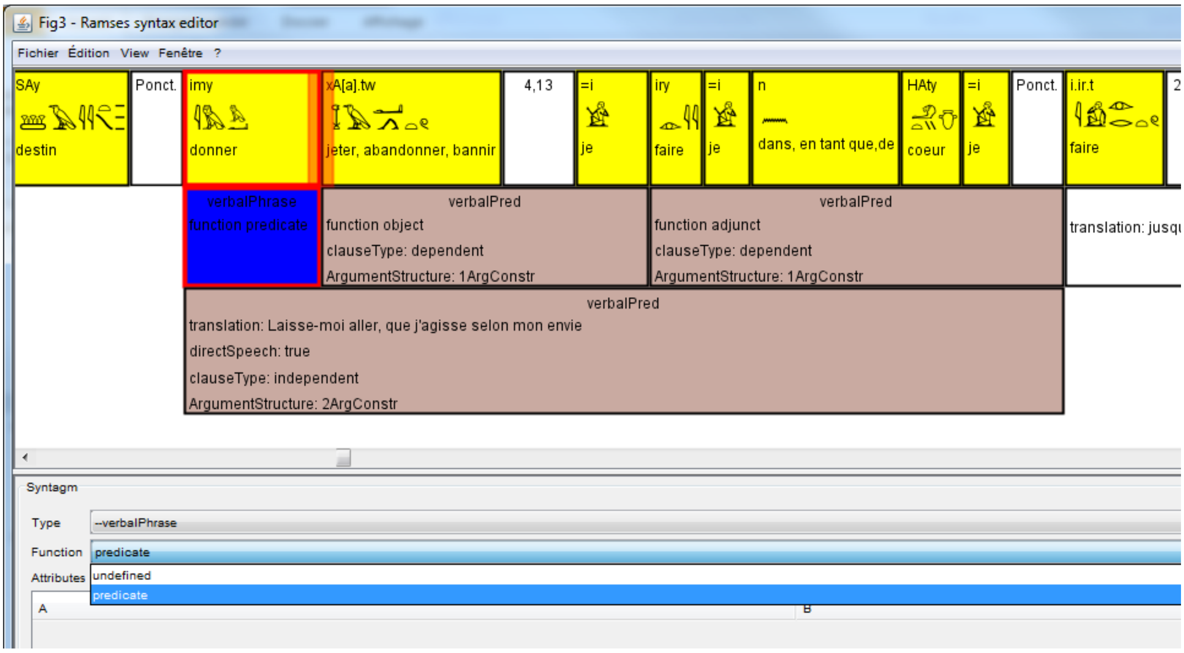
Attributes

A	B
translation	Elle accomplit les mois de grossesse
directSpeech	false
clauseType	unset
conversion	none
negation	false
interrogative	false
topic	none
ArgumentStructure	2ArgConstr

Figure 3. Example of simplified annotation

Fig. 3 shows the current interface of the SyntaxEditor when editing the sentence *skm=s ibd.w n msw* litt. "she passed the months of pregnancy" from the *Doomed Prince.* It features a number of interesting capabilities of the system. One can notice that the content of the noun phrase construction *ibd.w n msw.t* "months of pregnancy" has not yet been analyzed. This can be achieved at any point: it is possible to build a bottom-up analysis by grouping the words in larger constructions, and combining phrasal constructions into complex clausal constructions, then into whole sentences and possibly larger units. At the same time, a top-down analysis is also possible, first chunking the sentence-level constructions, then the main clausal and phrasal constructions, down to individual words. Actual practices of annotators are of course likely to mix the two approaches.

It is worth noticing however that the function of any element or construction directly depends on the larger construction to which it belongs: the annotation of the functions or roles of individual elements or constructions can only be achieved top-down. As shown Fig. 4, the function of the imperative *imy* "give" can be annotated only after the independent main clause type has been defined as a verbal predication and the only available option, in this case, is to annotate it with the "predicate" function.



verbalPhrase
function predicate

verbalPred
function object
clauseType: dependent
ArgumentStructure: 1ArgConstr

verbalPred
function adjunct
clauseType: dependent
ArgumentStructure: 1ArgConstr

translation: jusqu

verbalPred
translation: Laisse-moi aller, que j'agisse selon mon envie
directSpeech: true
clauseType: independent
ArgumentStructure: 2ArgConstr

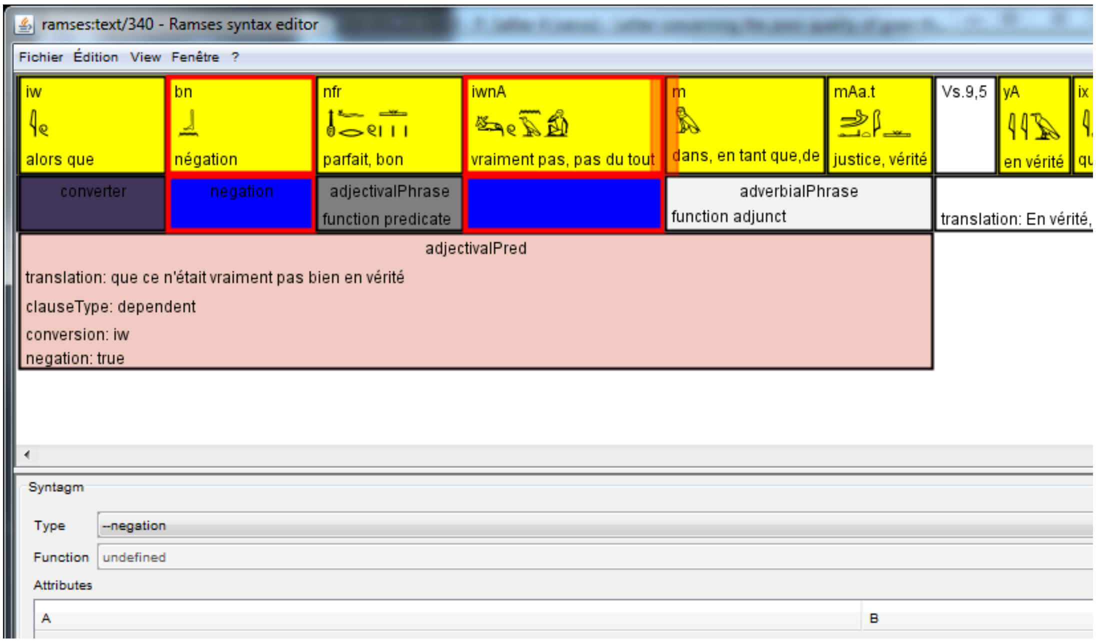
Syntagm
Type --verbalPhrase
Function predicate
Attributes undefined
predicate

Figure 4. Annotation of function for a verbal phrase

Coming back to Fig. 3, it also features a variety of attributes attached to the groups. The currently selected construction, a verbal predication, is outlined in red and specifies its children elements. The construction itself has attributes of different kinds: the translation is a simple free-text attribute ("elle accomplit les mois de grossesse"). Then, we have a number of boolean attributes (usually with three possible values, 'true', 'false'/'none', and 'unset'). And finally, the ArgumentStructure attribute has a value assigned from a list (here, '2ArgConstr').

In Fig. 5, one can see a discontinuous phrase.[21] The negation *bn* … *iwnꜣ* is considered as one element only. The groups are actually considered to be sets of elements, not spans of text (although, of course, this is the most frequent case).

[21] This example of discontinuous phrase is meant as an illustration and admittedly disputable; *bn* and *iwnꜣ* are — most probably — still two independent phrases in Late Egyptian, see Winand 1997.



iw — alors que — converter

bn — négation — negation

nfr — parfait, bon — adjectivalPhrase, function predicate

iwnA — vraiment pas, pas du tout

m — dans, en tant que,de

mAa.t — justice, vérité

adverbialPhrase, function adjunct

Vs.9,5

yA — en vérité

ix — qu

translation: En vérité,

adjectivalPred
translation: que ce n'était vraiment pas bien en vérité
clauseType: dependent
conversion: iw
negation: true

Syntagm
Type: --negation
Function: undefined
Attributes
A | B

Figure 5. Discontinuous negation

3. EVOLUTIONARY ANNOTATION SCHEME AND SYNTAX DESCRIPTION LANGUAGE

3.1. *A brief look in the rear-view mirror*

When the project was started in 2006, the immediate priority was to allow text encoding, even on an elementary level. Hence, the first versions of the software (a "TextEditor" coupled to a "Lexicon-Editor") were developed to allow the quick creation of a fully lemmatized corpus, with complete information about word spellings and inflexions, but without any syntactic grouping.

Due to the high variability of Late Egyptian orthography, and to the time-consuming nature of hieroglyphic text encoding, particular attention was paid to providing an efficient interface. As a result, we started our work by designing the lexicon structure and the corresponding software (LexiconEditor). The text database itself (TextEditor), containing the lemmatised texts, was created a little later.

The lexicon contains three kinds of entities: lemma, inflexions, and spellings. A lemma corresponds more or less to a main dictionary entry. The types of possible inflexions depend on grammatical categories or parts-of-speech that are defined at the level of the lemma; these inflexions are hierarchically subordinated to the lemmata in the LexiconEditor. For instance, we have a lemma for the verb *ḥtp*, "to be satisfied"; this lemma is attested in the corpus under different verbal inflexions (verbal morphology): e.g. infinitive, imperative, perfective, subjunctive, participle, etc. Due to the non-normative and defective nature of the Ancient Egyptian graphemic system, the spellings can be shared between one or more inflexions of different lemmata. For instance, the spelling *b3k* is connected with several inflections of the verb *b3k* "to work" (infinitive, old perfective, subjunctive and participle) as well as with the etymologically related substantive *b3kw* "task, labour".

The various problems we experienced when developing the LexiconEditor (and, to a lesser extent, the TextEditor) point out a number of desirable features for the kind of software we were developing. The main problem we faced is that, besides its general structure, such a database tends to change a lot in its early beginning with respect to its tagset (labels used, number and structure of the parts-of-speech tags, types and values of the attributes for annotating the lemmata and inflexions). Accordingly, the changes and evolutions needed ranged from small modifications in category labels to significant structural changes (removal or creation of new types of inflexions or parts-of-speech for example). Consequently, we had to write ad-hoc software on several occasions in order to carry out those modifications.

For the syntactic analysis, we wanted to avoid both falling into the same traps as well as writing the grammar before annotating the corpus, as this part of the project is intrinsically an experimental enterprise (we do not know for sure what categories will be used in the final version of the database). At the same time, we could take for granted that important changes would also occur in the syntactic tagset during the annotation of texts. Those changes would undoubtedly involve the addition of new analytical categories and the subtraction of some existing ones. Furthermore, when this occurs, the annotation on the previously analysed texts must not be lost, i.e. the software must be able to handle multiple annotation schemes.

For all those reasons, it was quite obvious that a great level of abstraction in software would be a major improvement for the SyntaxEditor annotation tool: the various categories which can be used in a syntactic analysis had to be explicitly defined in a self-standing annotation scheme that would allow the system users, i.e. the members of the Egyptological team, to perform changes within the tagset and basic grammar on their own — without any programming work — by using an intelligible Syntax Description Language (SDL) that they could easily modify and update.

3.2. *Control over the analysis structure*

Ramses is a collective, but centralized enterprise. Accordingly, it wouldn't be reasonable to allow each annotator to create the categories and attributes he needs on the fly. It would otherwise result in a complete inconsistency of the annotated data. Thus it is necessary to have some control over what constructions are available, and what attributes they can have. On the other hand, as stated in §3.1, it is equally inevitable that new categories will be needed at some point, and that significant changes, such as splitting a category in two, or merging two categories into one, will occur.

Faced with this problem, we decided to build on ideas present in Mazziota's *Notabene* (see 2010a and 2010b), and inspired by the semantic web, where the frame of each analysis is explicitly described. Instead of using Web Ontologies, we decided however, to create an ad-hoc language, specialised in the description of syntactic formalism, a Syntax Description Language.

3.3. *The Syntax Description Language (SDL)*

The formalism, built in the "annotation scheme", describes a kind of very simple "loose grammar" of the language, based on the basic principle that the grammar of any language is made up of taxonomic networks of families of constructions. The resulting description is saved in the database, and is identified by a name. An analysis is created using one particular annotation scheme (although it is possible to change it afterwards). Figure 6 illustrates a simplified and highly incomplete annotation system.

```
ANNOTATION SCHEME "Scheme_CxG_Test"

TYPE definiteness ENUM unset defined undefined doubtful ENDTYPE

GROUP construction
   ATTR comment TEXT * ENDATTR
ENDGROUP

// Phrases

GROUP phrasalConstr EXTENDS construction
ENDGROUP

GROUP nounPhrase EXTENDS phrasalConstr
   ATTR defined definiteness ONE unset ENDATTR
ENDGROUP
```

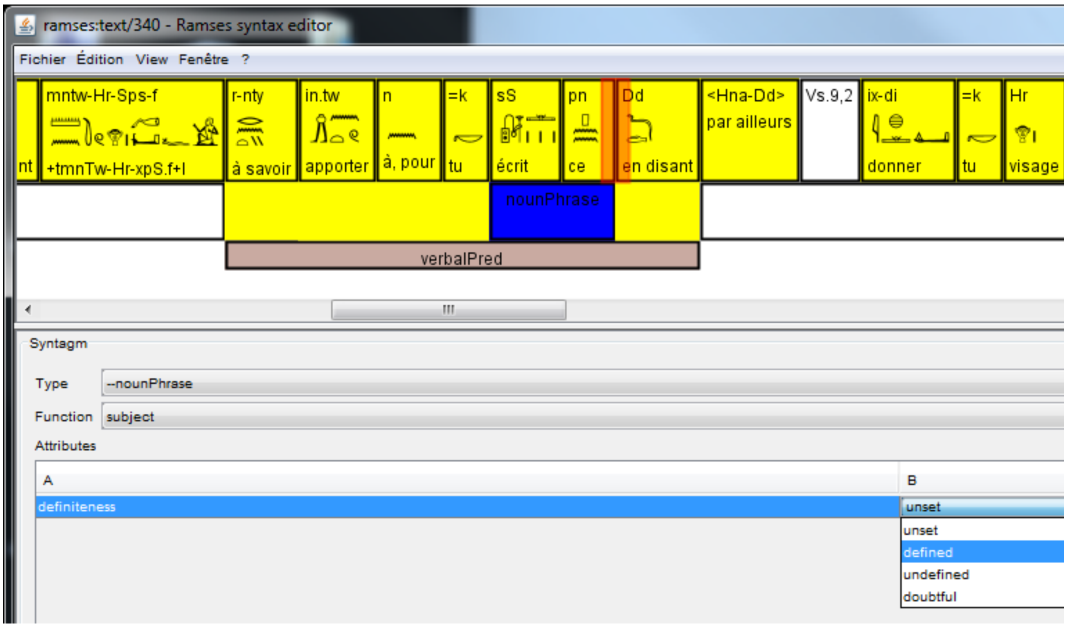
```
// Clauses
GROUP clausalConstr EXTENDS construction
   CHILD adjunct CHILDTYPE adverbialPhrase ENDCHILD
ENDGROUP
```

Figure 6. Simplified annotation scheme

Each annotation scheme name is given at the beginning of the file, in the present example `Scheme_CxG_Test` (it can be more or less anything).

Then, there are two kinds of elements in the SDL. First, one can describe `types`, which are used as attribute values. Here, we define the `definiteness` type, with four possible values: `unset`, `defined`, `undefined` and `doubtful`. The most important kind of element is the `group`. It describes what syntactic construction may exist in our description: any `group` is described by a name (which will be its label in the SyntaxEditor), and by its possible attributes. For instance, groups of type `nounPhrase` have an attribute called `defined`, of type `definiteness` (the `type` created just above). Other possible types are BOOLEAN for yes/no attributes, and TEXT for attributes whose value is a free text, like translations.

One can further specify whether an attribute is mandatory for a particular group, and indicate how many times (once or repeatedly) the attribute can be found. For instance, the `comment` attribute in `group` can be repeated (that is the meaning of the asterisk `*` character).



Syntagm

Type --nounPhrase

Function subject

Attributes

A	B
definiteness	unset

unset
defined
undefined
doubtful

Figure 7. Use of attributes

As it is likely that some constructions will have common attributes, we have implemented an “inheritance” system between groups. A group type can be declared to “extend” another group, in which case it will receive the same attributes (and possibly more). For instance, `clausalConstr` extends `construction` in Fig. 6, which means that all clausal constructions inherit the `comment` attribute defined for `group`. If we were to define sub-types of clauses, they would all receive the characteristics of `clausalConstr`, in addition to their own.

As exemplified in Fig. 7, the attributes definition is used to fill the list of possible attributes values in the SyntaxEditor.

3.4. *Children and relations*

In order to define explicitly the functions that any element can fulfil in the construction it belongs to, the SDL allows possible syntactic subordinates of a group to be declared as `children`. `Children` describe both a syntactic function (for instance, one of the children of a verbal predication can be its subject), and the kind(s) of group(s) which can fill this function.

A definition like:

```
GROUP monoclausalConstr EXTENDS clausalConstruction
  CHILD adjunct CHILDTYPE adverbialPhrase ENDCHILD
ENDGROUP
```

defines one kind of child for `monoclausalConstr` ("mono-clausal construction"): the adjunct, which can only be adverbial phrases in this example. When editing a group's attributes in the SyntaxEditor, a list proposes all the functions the group can have relative to its parent.

It is important to note that inheritance can be used here as well. It is possible to define verbal predicative patterns as follows:

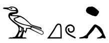
```
GROUP verbalPred EXTENDS monoclausalConstr
  CHILD subject CHILDTYPE nounPhrase ENDCHILD
  CHILD predicate CHILDTYPE verbalPhrase ENDCHILD
ENDGROUP
```

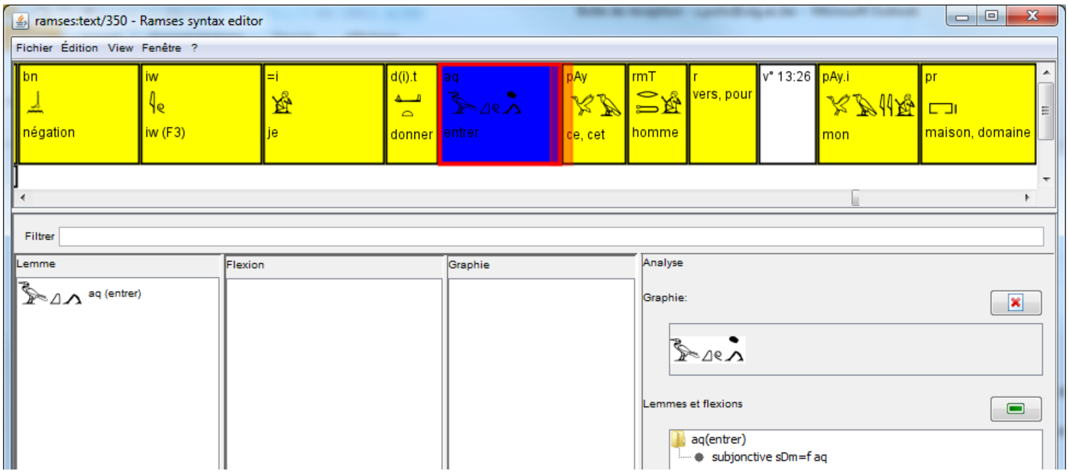
In this case, the verbal predicative pattern is defined as a kind of mono-clausal construction and can have three kinds of children: adjuncts, inherited from mono-clausal construction, and a subject as well as a predicate, which are specific to the verbal predicative pattern.

In the present state of the SDL, an element has only one parent, which means that the syntactic analysis is practically limited to trees. Now, as was mentioned in the introduction, there are many relationships between phrases that are not of a strictly hierarchical nature. One can think of anaphoric relations as a typical example. (In this domain, there is also an interest among the members of the Ramses project in the interaction between narration and discourse in the texts: we want to be able to link related instances of indirect speech, for example). For all those reasons, we need a second type of link between groups that enables us to annotate graph structures, and we have termed this link `relation`. Relations have the same kind of definition as children do; this time, however, there is no embedding constraint at all. At the moment, there is no support in the SyntaxEditor for creating actual relations, but they will probably be materialised by arrows between the two related groups.

4. ERGONOMICS OF THE USER INTERFACE

Considering that a large part of the treebank will be created manually in a first step, we needed a powerful annotation tool: this user interface, called "SyntaxEditor", had to follow some ergonomic principles in order to minimize the need for training and speed up the annotation work. In this section, we will briefly describe the steps for manually annotating with syntactic structures an (admittedly quite simple) independent clause in a text; thus we illustrate further (see Fig. 3-5; 7) the capabilities of the SyntaxEditor.

The text to be annotated is first imported into the SyntaxEditor as a sequence of tokens. Fig. 8 shows clearly that the other levels of annotation are easily accessible in the SyntaxEditor, simply by clicking on the token one wishes to have information about; here, the verb *ꜥḳ* "to enter" has a spelling  and a defined inflexion (subjunctive).



Lemme: aq (entrer)

Lemmes et flexions: aq(entrer)
subjonctive sDm=f aq

Figure 8. Other levels of annotation in the SyntaxEditor: spelling, lemma and inflexion

In a second step, the annotator selects the elements to be grouped together or the elements to be chunked in the clause; both bottom-up and top-down approaches are possible. The annotator can define the construction types either directly or afterwards. The SyntaxEditor allows for much flexibility in this respect: as shown in Fig. 9, the main constructions of the sentence have been identified, but only two of them actually received a type; the other ones have simply been tagged as generic groups.

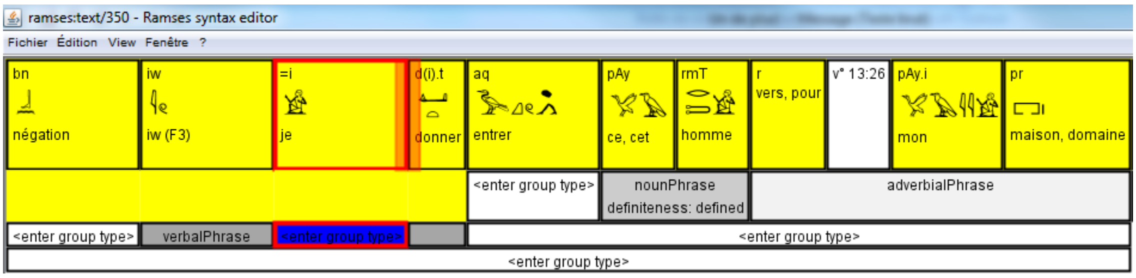
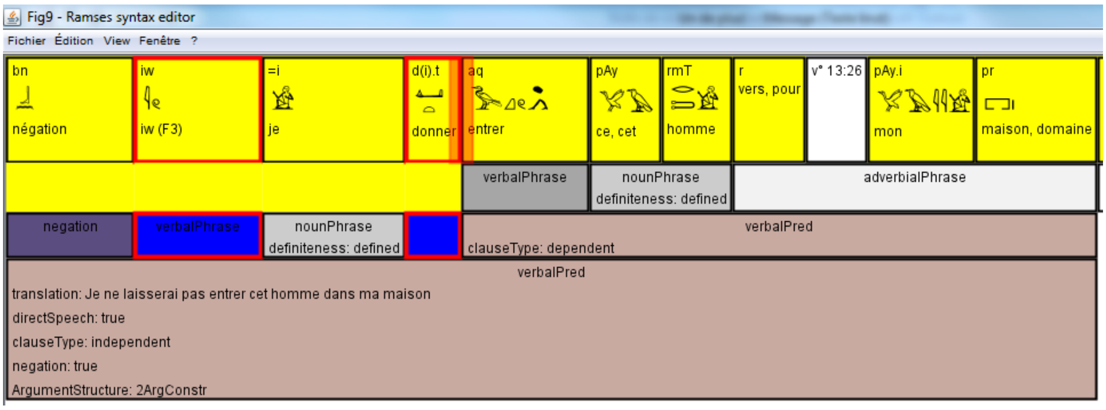


Figure 9. Creation of groups and identification of group types

Once all the top-level constructions have been annotated with types and related attributes, the syntactic analysis appears as in Fig. 10. Two points are worth mentioning here: (a) if the value of the attribute of any construction is the default one, it is not actually displayed in the box summarizing the data;[22] (b) in order to avoid duplicating any kind of information, the verbal phrase *iw* … *di.t* "(I) will cause" is analysed as a single discontinuous group — the information about the status of the conjugation base *iw* and the infinitive *di.t* are already annotated at the level of the TextEditor.

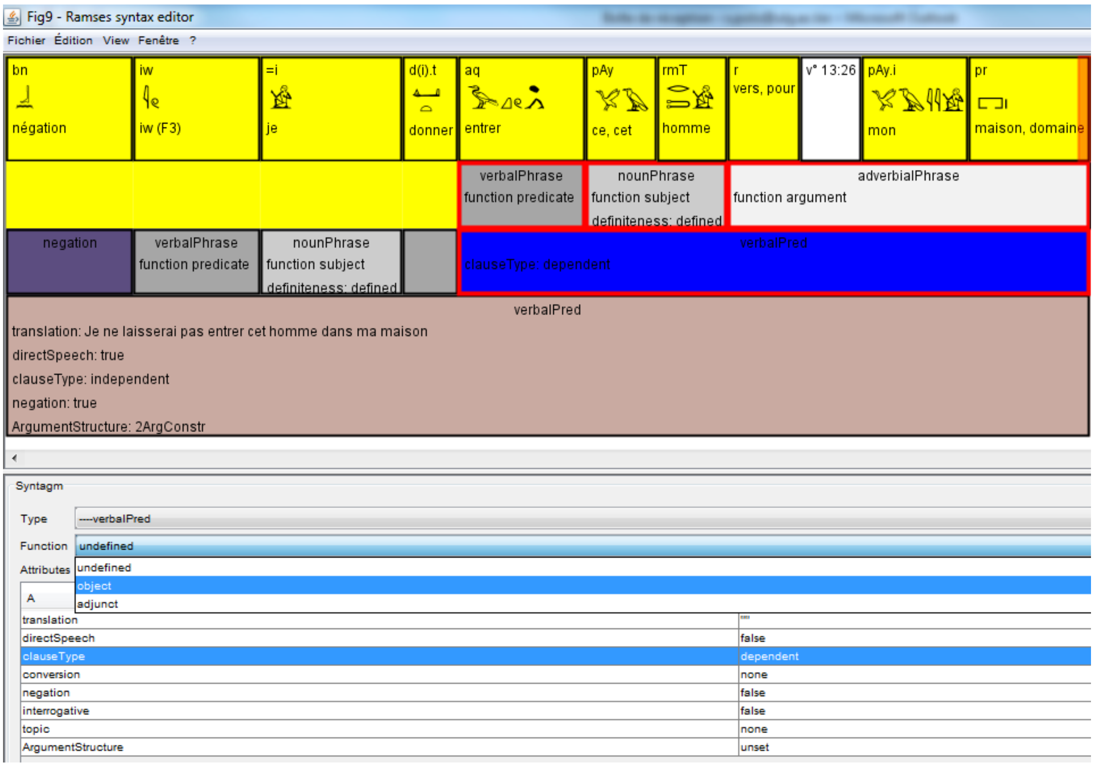


translation: Je ne laisserai pas entrer cet homme dans ma maison
directSpeech: true
clauseType: independent
negation: true
ArgumentStructure: 2ArgConstr

Figure 10. A clause with the top-level construction identified

[22] In the example of Fig. 10, there is no data regarding the interrogative status of the main clause construction, for its default value is "false".

The third step consists of the annotation of the functions of the constructions. As already explained (see §3.4), this step can only be achieved when the status of the higher level construction have been defined, for the only available functions in the SyntaxEditor for a given element are the ones that are defined to be acceptable for this type of element in a higher construction within the annotation scheme. In Fig. 11, we see that an unconverted dependent clause can only be annotated as a complement clause or as an adjunctive (consecutive) clause when it is an element of an independent main clause.



bn	iw	=i	d(i).t	aq	pAy	rmT	r	v° 13:26	pAy.i	pr
négation	iw (F3)	je	donner	entrer	ce, cet	homme	vers, pour		mon	maison, domaine

verbalPhrase
function predicate

nounPhrase
function subject
definiteness: defined

adverbialPhrase
function argument

negation

verbalPhrase
function predicate

nounPhrase
function subject
definiteness: defined

verbalPred
clauseType: dependent

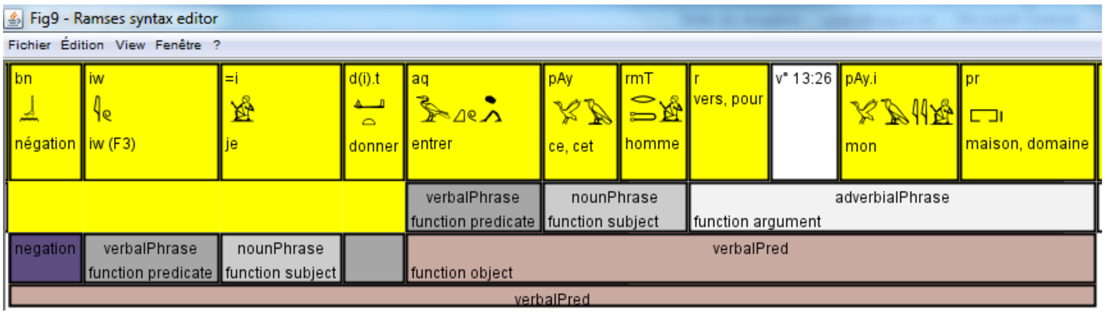
verbalPred
translation: Je ne laisserai pas entrer cet homme dans ma maison
directSpeech: true
clauseType: independent
negation: true
ArgumentStructure: 2ArgConstr

Syntagm
Type ----verbalPred
Function undefined
Attributes
undefined
object
adjunct

A	
translation	""
directSpeech	false
clauseType	dependent
conversion	none
negation	false
interrogative	false
topic	none
ArgumentStructure	unset

Figure 11. Annotation of syntactic roles

Finally, as illustrated in Fig. 12, the constructions can be flattened using a toggle function in order to limit the information to the types and functions, without displaying the data that concerns the attributes.



bn	iw	=i	d(i).t	aq	pAy	rmT	r	v° 13:26	pAy.i	pr
négation	iw (F3)	je	donner	entrer	ce, cet	homme	vers, pour		mon	maison, domaine

verbalPhrase
function predicate

nounPhrase
function subject

adverbialPhrase
function argument

negation

verbalPhrase
function predicate

nounPhrase
function subject

verbalPred
function object

verbalPred

Figure 12. Annotation of syntactic roles

5. FUTURE DEVELOPMENTS

There are many points still on the to-do list for the SyntaxEditor. Some of them are rather mundane, but very important, such as the implementation of a versioning system, in order to track changes in the texts.

Other are more exciting, mostly those that include a natural language processing component. There are indeed several areas where the system should be cleverer. First, it would be important to be able to define, in a reasonably simple way, validation rules, to detect problems in the analysis. Another important area of work is the developments of tools for updating "old" analysis: when a new annotation scheme becomes available, we will certainly want to transfer the previous work done with the old annotation scheme.[23] In some cases, it will be easy a task (e.g. when a category is common to both schemes), but some changes will be trickier, and others will require some manual editing (an important feature of the current software is that the old annotations are kept when changing the annotation scheme for a text).

Finally two more ambitious tasks are the implementation of a search engine and the development of a parser for Late Egyptian texts.

5.1. *Search engine*

Such a syntactically annotated corpus would evidently be useless without efficient search facilities. There are a number of existing tools which are suitable for searches in structured text databases. To name but a few: emdros, TIGERSearch, and the Intex family (Intex, Unitex and Nooj). Given the current state of Ramses, we will likely base our search engine on TIGERSearch,[24] which can be easily interfaced with Ramses as both are written in Java. It satisfies two conflicting requirements: it is easy to learn (close to grammar formalism) and its expressiveness is constrained in order to guarantee efficient query processing. It is of course too early to describe the engine, but it is possible to give an idea of the possibilities of existing systems. For instance, given a correctly structured database, TIGERSearch would allow queries like:

```
(#n1:[cat=”nounPhrase” & defined=”undefined”]
.*
#n2:[cat=”relativeClause”])
& (#n2 >~antecedent #n1)
```

which would find all relative clauses whose antecedent is not defined. Note that such a query cannot be done on a corpus annotated only for part-of-speech and lemma. Syntactic annotation is definitely needed. TIGERSearch also provides a graphical interface to avoid the direct use of the query language, and probably makes simple queries easier to write.

5.2. *Interactive parser*

It has already been shown[25] that noun phrase constructions (at least the ones that do not include relative or participial clauses) can be parsed with a very high degree of accuracy using automata, which explains why we decided to focus first on the higher rank constructions (and not on the phrasal constructions).

Now, the most ambitious work in front of us in this domain is certainly the partial automation of the analysis, which is part of the PhD thesis on which Benjamin Martin Leon (University of Liège) is working. Much like the parser which has been developed in the framework of the TIGER Treebank,[26] we include in our future plans an "interactive parser": "[i]nteractive annotation is an efficient combination of automatic parsing and human annotation. Instead of having an automatic parser as preprocessor and a human annotator as postprocessor, the two steps are interwoven in our approach." Moreover, given the length of attestation of Late Egyptian linguistic data (more than five hundred years), the wealth of registers and the strong influence of the writing support on the spellings, the

[23] In this respect, see the comparative procedure in Mazziotta (2010b: §2.3).

[24] For the query language, see König & Lezius 2002a & 2002b.

[25] See Benjamin Martin Leon's unpublished MA thesis: "Projet Ramsès: réalisation d'une bibliothèque de traitement à états finis" (Liège; 2008-2009).

[26] See Brants *et al.* 2002: §3.1.

parser will have to take into account mark-up data on the corpus if one wishes to reach an acceptable degree of accuracy.

6. Conclusion

The SyntaxEditor inaugurates the next important step in the Ramses project. The decision — taken, to be honest, more out of necessity than on theoretical grounds — to start with a "simple" lemmatised corpus now allows us to start working on this level of annotation with a large database already tagged for lemma, inflexions and spelling. We envisage it as a solid foundation on which to build this next stage. In particular, having all these data means that we can test syntactic hypotheses more easily, and try statistical methods if needed.

From a linguistic viewpoint, we take seriously the assumption of Construction Grammar that *constructions* are the basic units of syntactic representation; consequently, 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.

Bibliography

BAMMAN, David & Gregory CRANE. 2011. The Ancient Greek and Latin Dependency Treebanks, in: Caroline SPORLEDER, Antal VAN DEN BOSCH & Kalliopi ZERVANOU (eds.), *Language Technology for Cultural Heritage*, Springer, Foundations of Human Language Processing and Technology. http://nlp.perseus.tufts.edu/syntax/treebank/publications.html

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

BLACHE, Philippe. 2000. À quoi sert l'annotation syntaxique de corpus, in: BILGER (ed.), p. 82-94.

BRANTS, Sabine, Stefanie DIPPER, Silvia HANSEN, Wolfgang LEZIUS, George SMITH. 2002. The TIGER Treebank, in: *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, p. 24-41.

CROFT, William. 2001. *Radical Construction Grammar. Syntactic Theory in Typological Perspective*, Oxford, Oxford University Press.

DIPPER, Stefanie & Michael GÖTZE. 2005. Accessing heterogeneous linguistic data — Generic XML-based representation and flexible visualization, in: *Proceedings of the 2nd Language & Technology Conference*, Poznan, p. 206-210.

GOLDBERG, Adele E. 1995. *Constructions. A Construction Grammar Approach to Argument Structure*, Chicago, The university of Chicago Press.

—. 2006. *Constructions at Work. The Nature of Generalization in Language*, Oxford, Oxford University Press.

HAJIC, Jan. 1999. Building a syntactically annotated corpus: The Prague Dependency Treebank, in: Eva HAJIČOVÁ (ed.), *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevova*, Prague, Charles University Press, p. 106-132.

HAUG, Dag *et al.* 2009. Computational and linguistic issues in designing a syntactically annotated parallel corpus of Indo-European languages", in: *Traitement Automatique des Langues* 50/2, p. 17-45.

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

KÖNIG, Esther & Wolfgang LEZIUS. 2002a. *The TIGER language — A Description Language for Syntax Graphs. Part 1: User's Guidelines*, University of Stuttgart. (http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/papers/tigerLanguage.ps.gz)

—. 2002b. *The TIGER language — A Description Language for Syntax Graphs. Part 2: Formal Definition*, University of Stuttgart. http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/papers/tigerLang Form.ps.gz

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

MARCUS, Mitchell P., Beatrice SANTORINI & Mary Ann MARCINKIEWICZ. 1993. Building a large annotated corpus of English: The Penn Treebank, in: *Computational Linguistics* 19/2, p. 313-330.

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. Building 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, p. 142-146.

MCGILLIVRAY, Barbara, Marco PASSAROTTI & Ruffolo PAOLO. 2009. The *Index Thomisticus* Treebank project: Annotation, parsing and valency lexicon, in: *Traitement automatique des Langues* 50/2, p. 103-127.

MEL'ČUK, Igor. 1988. *Dependency Syntax: Theory and Practice*, University of New York Press, Albany.

PETERSEN, Ulrik. 2004. Emdros — A Text Database Engine for Analyzed or Annotated Text, in: *ACL, COLING 2004 Geneva, 20th International Conference on Computational Linguistics*, Volume 2, p. 1190-1193. [The application is available at http://emdros.org/]

POLGUÈRE, Alain & Igor MEL'ČUK (eds.). 2009. *Dependency in linguistic description*, Amsterdam-Philadelphia, John Benjamins.

POLIS, Stéphane, Anne-Claude HONNAY & Jean WINAND. Current volume. Building an annotated corpus of Late Egyptian. The Ramses Project: Review and perspectives.

RECASENS, Marta, M. Antònia MARTÍ & Mariona TAULÉ. 2007. Where anaphora and coreference meet. Annotation in the CESS-ECE corpus, in: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2007)*, Borovets.

ROBINSON, Jane. 1970. Dependency structure and transformational rules, in: *Language* 46, p. 259-285.

STEDE, Manfred. 2004. The Potsdam Commentary Corpus, in: *Proceeding DiscAnnotation '04. Proceedings of the 2004 ACL Workshop on Discourse Annotation*, p. 96-102.

TESNIÈRE, Lucien. 1965. *Éléments de syntaxe structurale*, Paris, Klincksieck.

WINAND, Jean. 1997. La négation *bn … ỉwnꜣ* en néo-égyptien, in: *Lingua Aegyptia* 5, 223-236.

Automated Text Categorization in a Dead Language[*]

The Detection of Genres in Late Egyptian

Stéphanie GOHY[§], Benjamin MARTIN LEON & Stéphane POLIS[§]

F.R.S.-FNRS[§] – Université de Liège

1. INTRODUCTION

Automated Text Categorization (ATC) is common in all applicative domains that involve information retrieval, organization and management.[1] It can be defined as the "activity of automatically building, by means of machine learning (ML), *automatic text classifiers*, i.e. programs capable of labeling natural language texts from a domain D with thematic categories from a predefined set $C = \{c_1,\ldots,c_{|C|}\}$" (Sebastiani 2002: 3; Debole & Sebastiani 2004: 81). The ever-growing quantity of textual material available online and the correlative extension of applicative contexts led ATC to become one of the major subfields of information system research;[2] accordingly, the last twenty years have seen the development of innovative approaches to the inductive construction of text classifiers.

Automatic Genre Identification[3] (AGI), the topic of the present study, is one particular subfield of ATC. With AGI, the categories (i.e. the textual genres) are predefined — one speaks of a "supervised learning method"[4] — and each text is assigned by the classification method (called the "classifier") to one of these categories[5] — one speaks of "non-overlapping categories" or "single-label classification scheme".

Unlike AGI of web documents[6] or AGI applied to large-scale modern corpora, AGI in a dead language with a corpus of limited size is not primarily directed towards applications such as text filtering, document organization or word-sense disambiguation. Instead, the aim of AGI in a language like Late Egyptian is two-fold:

[*] We are grateful to Todd Gillen and Eitan Grossman for comments on first drafts of this paper.

[1] ATC is especially important in library sciences, in media (e.g. with topic spotting and content sorting of news feeds from press agencies) and, more generally, on the web, where the many applications of ATC range from web page classification (which allows structured browsing, see below) to spam filtering. For a convenient introduction to machine learning approaches to text categorization, see Basili & Moschitti 2005.

[2] A bibliography on the topic (updated until 2007) is available online at http://liinwww.ira.uka.de/bibliography/Ai/automated.text.categorization.html.

[3] See the special issue of the *Journal for Language Technology and Computational Linguistics* devoted to this topic (Santini *et al.* 2009, with previous literature).

[4] As such, it differs from *(hierarchical) text clustering*, an "unsupervised method" that aims at automatically grouping documents together in a set of categories that is *not* predefined.

[5] In other ATC domains, texts can belong to several overlapping categories. The same holds for AGI when applied to web pages, where multi-label approaches are becoming more and more common, see e.g. Vidulin *et al.* 2009.

[6] Where this procedure can facilitate the access to appropriate results of search engines; see e.g. Lim, Lee & Kim 2005, with previous literature.

(1) From a linguistic point of view, we aim at describing the norms of the register(s) that are characteristic of genres. AGI works here as a heuristic tool: different features can be taken into account for AGI and the relevance of each feature for describing a register can be evaluated based on the performance of the classifiers using this feature. It means that each text, in its singularity, can be compared against generalizations about the linguistic norms of genres that are learned — automatically and inductively — based on selected linguistic features of a training set. The ultimate research goal is the study of the relation between registers, genres and discourse types at a linguistic level.

(2) From a more practical viewpoint, AGI will help to enhance the performance of Natural Language Processing (NLP) tools currently under development for Late Egyptian in the framework of the Ramses project (see §2): as pointed out by several scholars,[7] the performance of taggers and parsers can be significantly enhanced once the genre of a text is known.

The structure of the paper is as follows. In the next section (§2), we briefly present the Ramses corpus and we introduce the levels of annotation integrated therein, thereby clarifying which linguistic features can be used for AGI in Late Egyptian. The following section (§3) is devoted to a survey of the types of features that one finds in the literature for AGI; given the abovementioned goals, the present study will mainly refer to linguistic criteria. In the last section (§4), we apply three 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 and we test their respective performance with lexical and morphological features.

2. THE RAMSES ANNOTATED CORPUS OF LATE EGYPTIAN TEXTS

Ramses is a manually annotated corpus of Late Egyptian texts currently under construction at the University of Liège.[8] This corpus will ultimately include all extant Late Egyptian texts and, more broadly, all the written sources whose linguistic registers attest Late Egyptian linguistic features from the 18th dynasty down to the Third Intermediate Period (ca. 1350-700 BCE). The size of the corpus is estimated to ca. 1 million words on completion, and consists of ca. 300 000 tokens as of late 2011. The Ramses corpus is annotated for lemmata[9] and inflexions (see Fig. 1). The syntactic layer is still in its test phase.[10] Additionally, the corpus includes a graphemic level (hieroglyphic spellings are fully supported) and corpus mark-up (i.e. a set of metadata about date, nature of the writing support, writing system, place of origin, etc.).

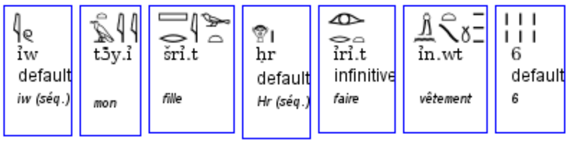


Figure 1. A sentence in Ramses' TextEditor

The tests for AGI in Late Egyptian have been performed on 322 texts belonging to seven genres that differ quite significantly from one another: letters (LET.), judicial documents (JUD.), oracular questions (OR.), educational texts (EDU.), monumental texts (MON.), hymns and prayers (HYM.), and administrative texts (ADM.).

[7] See e.g. Kessler *et al.* 1997.

[8] See Polis, Honnay & Winand (current volume), with previous literature.

[9] Lemmata are tagged with information on part-of-speech, animacy, and basic semantic class.

[10] See Polis & Rosmorduc (current volume).

It should be stressed that the distribution of the texts between these categories (as well as their respective length in terms of tokens) is highly unbalanced. Here follows a list of the number of texts for each genre (arranged by quantity of attestation):

142	letters,
47	judicial documents,
41	oracular questions,
29	educational texts,
28	monumental texts (royal),
20	hymns,
15	administrative texts.

Given the small size of the training set of documents for some categories, we expected the performance of AGI to vary significantly between genres.

3. WHAT ARE THE FEATURES USED IN AUTOMATIC GENRE IDENTIFICATION?

Ever since Aristotle's *Poetics*, discussions about the principles at stakes for classifying literary texts, at first, and, subsequently, any type of written production, did not lead to a broad consensus of opinion between scholars. The main reason for this is most certainly that genres are embedded in complex socio-cultural practices (genres are "social institutions") and span a wide variety of communicative situations and functions. When talking of genres, one is dealing with a protean concept that appeals to various strata of analysis and, consequently, relies on heterogeneous classificatory principles: in any human society, many parameters can be taken into consideration for classifying textual material.

The approach of AGI, which already has a long history in computational linguistics,[11] is empirical rather than theoretical: it has been trying to reach the best performance in classification by testing empirically what kinds of parameters or features produce the best results. As Lim *et al.* (2005: 1264) put it, "selecting features that can make a clear distinction among the genres is the core of automatic genre classification." Four main types of textual features,[12] which correlate with observable surface cues, have been used in the body of literature on AGI:[13]

(1) **Material features**, which can be extracted from corpus mark-up, such as date of composition, communication medium, type of writing, place of origin.

(2) **Structural features.** These have to do with the types of formatting device (the presence of headings, of lists, of different typefaces, of images, etc.) as well as with other formal properties of the texts (e.g. the number of paragraphs, sentences, words, characters in a text; the number of words or character per sentence, etc.).

[11] See Biber's pioneering work (1986; 1988, 1993a; 1993b, 1995) on genre variation. Biber aims at building *inductively* a typology of texts based on "dimensions" of genre variation. In a nutshell, his method consists in analyzing the quantitative distribution of numerous co-occurring linguistic features that are considered to be characteristic of one particular dimension (informational vs. involved; narrative vs. non-narrative; etc.). The main idea is that it is the co-occurrence of sets of markers that matters, rather than individual features in isolation. It is the combination of dimensions that defines genres. Each isolated feature (such as word length, type-token ratio, 2nd person pronouns, past tense verbs, phrasal coordination, conditional subordination, agentless passive, etc.), on the other hand, can be associated with several dimensions of variation. It is worth noticing that the types of linguistic features acknowledged by Biber have been very influential for later approaches to AGI.

[12] We restrict ourselves to features that are appropriate for written/printed material. Regarding those that are relevant specifically for web documents, see Karlgren *et al.* 1998; Lee & Myaeng 2002; Lim *et al.* 2005.

[13] Recent contributions in the field include Karlgren & Cutting 1994; Kessler *et al.* 1997; Karlgren *et al.* 1998; Michos *et al.* 1996; Stamatatos *et al.* 2000a; 2000b; Lee & Myaeng 2002; Malrieu & Rastier 2002; Jebari 2009; Santini *et al.* 2009.

(3) **Semantic features.** These are usually extracted based on the lemmata of texts that are taken to be indicative of their thematic contents. One remark is warranted here: if propositional content has been the focus of most ATC research, the assumption that text belonging to the same genre share similar semantic features is overly simplistic.[14] In practice, however, basic lexical counts — such as, for instance, the bag-of-words model,[15] whether or not weighted with statistical methods such as *tf-idf*[16] — have proven to perform relatively well in AGI.

(4) **Morpho-syntactic features.**[17] The broadly functionalist idea here (stressed e.g. by Biber in relation to genre variation) is that texts sharing similar communicative functions will use similar linguistic forms in order to fulfill these functions. At the syntactic level,[18] one finds in the literature features such as the proportion of nominalizations or topicalized sentences (to name but two), but also basic syntactic counts (like the number of words in a Noun Phrase; the ratio between the number of NPs and the total number of chunks; the average length of a parsed tree, etc.). At the morphological level, one can cite parts-of-speech related counts (such as the proportion of adverbs, nouns or pronouns, verb-noun ratio, etc.) and inflectional counts (number of passives, etc.).

In most scholarly works — given that the performance of the classifier is the main (or sole) goal — many features[19] belonging to these four categories are taken into account simultaneously, based on the assumption that any text "can be described in terms of an indefinitely large number of facets" or features (Kessler *et al.* 1997: 33).

The orientation of this study is quite different in this respect (see §1), since we use AGI mostly as a heuristic tool in order to identify the types of features that are relevant for the description of registers in Late Egyptian. Therefore, for this first application of AGI methods to Egyptian, we decided to exclude material and structural features and to test independently semantic and morphological features by focusing on the lemmata and inflexions that are characteristic of genres.[20]

4. MACHINE LEARNING METHODS FOR AUTOMATIC GENRE IDENTIFICATION: THREE CASE STUDIES IN LATE EGYPTIAN

In the field of AGI, the algorithms of classification are usually based on machine learning techniques: "a general inductive process automatically builds a classifier by learning, from a set of previously classified documents, the characteristics of one or more categories" (Sebastiani 2002: 1). The most

[14] On subject-classified vs. genre-classified data, see Lee & Myaeng 2002.

[15] This label refers to the fact that texts are envisioned as collections of words, with no attention to the order of words in texts or to their inflectional patterns. Other related methods are, for instance: most common word frequencies (coming from authorship attribution studies); the presence vs. absence of specific words as indicative of a genre; the vocabulary richness with type-token ratio V/N, etc.

[16] On the *tf-idf* weighting function, see Salton & Buckley 1988. "This function encodes the intuition that (i) the more often a term occurs in a document the more it is representative of its content, and (ii) the more documents the term occurs in, the less discriminating it is. […] This formula […] weights the importance of a term to a document in terms of occurrence consideration only, thereby deeming of null importance the order in which the terms occur and the syntactic role they play." (Sebastiani 2002: 14).

[17] See especially Beauvisage 2001.

[18] A special type of "syntactic" feature is related to punctuation cues and other delimiters (with counts of question mark, exclamation marks, etc.).

[19] See for example the 100 linguistically- and functionally-motivated features (or facets) taken into account by Santini (2010: 125).

[20] For the purpose of the present paper, the problem of the link between linguistic registers and genres has been significantly simplified, since we consider here the relation between genres and registers as a one-to-one relationship: each genre is linked to one and only one register, and conversely. Furthermore, we envision texts as units, putting no statistical overload on sentences possibly more representative of a genre; on this technique, see e.g. Ko *et al.* 2002; 2004.

frequent ways of building classifiers[21] are: probabilistic classifiers (based on Bayes' theorem), the Rocchio algorithm, the *k*-Nearest Neighbor algorithm, Decision Rule, Decision Trees, Neural Network classifiers, and Support Vector Machine method.

For the following case studies, we used three supervised learning methods called respectively (§4.1) the probabilistic *naïve Bayes classifier*, (§4.2) the *Support Vector Machine* and (§4.3) the less widespread *Segment and Combine* method.

The procedure followed with these three supervised learning methods is the following: 70% of previously classified texts have been used as *training set* in order to generate the prediction function using the learning algorithms. The remaining 30% of texts comprise the *test set*, which has been submitted to the prediction function in order to get predictions on the genres. Fig. 2 summarizes this procedure.

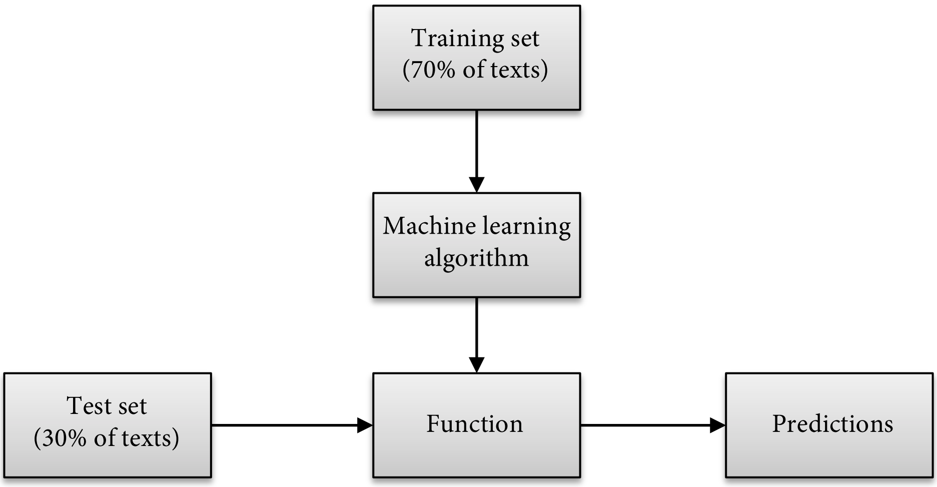


Figure 2. The learning and testing phases

4.1. *Naïve Bayes classifier*

In a first step, we consider the performance of one of the 'golden oldies' of classification methods, namely the naïve Bayes classifier. A naïve Bayes classifier is a simple probabilistic classifier[22] that computes the probability that a text belongs to different categories (in the present case, genres) based on Bayes' theorem. The texts are assigned to the category (genre) that received the highest probability.

The classifier is said to be *naïve* because of a strong independence assumption:[23] it makes the naïve hypothesis that features (namely words) are independent of each other (bag-of-words approach to document representation). Nevertheless, as is widely acknowledged in the literature on Information Retrieval, this very simple representation of texts has proven to be as effective as others.

The mathematical expression of the classifier reads as follows:

$$Genre(T)= \underset{G}{argmax}\ p(G)\prod_{i=1}^{n} p(W_i|G)$$

Figure 3. The naïve Bayes classifier for AGI

[21] See e.g. Sebastiani 2002: 24-40; Jebari 2009: 76-77.

[22] For a review of the uses of the naïve Bayes classifier, see Lewis 1998.

[23] This assumption makes the computation of the naïve Bayes classifier much more efficient than the exponential complexity of a pure Bayesian approach.

This formula says that the genre of a text *T* is the genre *G* for which the product between the prior probability of *G* and the conditional probabilities of each word W_i of *T* given *G* is maximal.

4.1.1. Performance of the naïve Bayes classifier

The tests performed with the naïve Bayes classifier give a global performance of 84.3% of texts that are well categorized. The confusion matrix[24] in Fig. 4 shows the details of the classification accuracy, genre by genre:

	LET.	JUD.	OR.	EDU.	MON.	HYM.	ADM.	**PERF.(%)**
LET.	138	2	0	0	0	0	2	**97.2**
JUD.	2	39	0	0	2	0	4	**83.0**
OR.	9	2	27	0	1	0	2	**65.9**
EDU.	5	0	0	21	1	1	1	**72.4**
MON.	0	0	0	0	28	0	0	**100.0**
HYM.	0	0	0	2	1	17	0	**85.0**
ADM.	0	2	0	0	0	0	13	**86.7**

Figure 4. Confusion matrix with the simple naïve Bayes classifier

This performance is quite good, given that the size of the training sets is small and that we are dealing with seven non-overlapping categories simultaneously. The details of this confusion matrix call for several remarks:

(1) It is noticeable that the texts of two genres are especially well classified: the monumental royal texts (100%) and the letters (97,2%). The reason for the very good performance of these two categories of texts is certainly twofold: the register of royal monumental texts is very high on the formality scale and highly standardized — it emulates the language of the past in many respects —, which probably set these categories quite clearly apart from the other genres of the corpus. In the case of letters, on the other hand, the good performance should certainly be related to the bigger size of the training set: this category is quantitatively larger than the others, which naturally leads to a better categorization (see §2).

(2) The misclassification of letters, judicial and administrative texts is revealing and very interesting from a linguistic point of view. Indeed, except for two cases, the texts belonging to these categories can be confused with each other, but not with other genres. This corroborates the intuition that the texts belonging to these genres use registers that are similar (and probably the closest to the Late Egyptian vernacular).

(3) Hymns and prayers, on the other hand, when misclassified, are categorized with texts belonging to the higher part of the formality scale, which also meets the linguist's intuition about the language of these texts.

(4) Finally, one should notice that the performance of the genre "oracular questions" is not especially good when compared with other genres (65,9%). This is most certainly due to the

[24] A confusion matrix is a table layout typically used with supervised learning methods. It allows the visualization of the performance of an algorithm. Each row of the matrix represents the instances of texts in an actual class, while each column represents the texts in a predicted class. The well classified texts are the ones belonging to the diagonal of the matrix. The last column, which does not belong to the matrix strictly speaking, gives the performance of each literary genre. The phrase "performance of a genre" is used here as a shortened form of "the performance of the classification system in correctly predicting a genre".

fact that the texts belonging to this category are very short and not very well differentiated from a thematic point of view: they most frequently consist of one or two short sentences written on limestone that were submitted to the divinity during his procession in order to get his opinion on daily life matters. The text in Fig. 5 is a typical example that reads *ns-sw Bꜣsꜣ* “does it belong to Bes?”

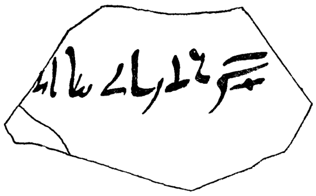


Figure 5. A typical oracular question (O. IFAO 866)

4.1.2. Integrating structural features

In order to enhance the performance of the naïve Bayes classifier for the development of NLP tools within the project (see under §2), we modified its mathematical expression so as to take into account the size of the texts in the corpus. The new classifier expression in Fig. 6 is the same as that of Fig. 3, except for the division by the difference (σ) between the number of words contained in a text (*T*) and the average number of words contained in texts belonging to the genre (*G*):

$$Genre(T)=\frac{\underset{G}{argmax}\; p(G)\prod_{i=1}^{n} p(W_i|G)}{|\sigma_{TG}|}$$

Figure 6. The naïve Bayes classifier modified to account for text length

This new expression of the classifier increases global performance by about 3%. Interestingly enough, the performance of oracular questions increases by almost 30%. One can further notice that, while the performance of other genres is approximately the same, that of educational texts increases and that of hymns decreases.

4.2. *Support Vector Machines*

Support vector machines (SVM) are universal learning algorithms earners used to solve classification and regression problems. They were introduced by Vapnik in 1979 (see Vapnik 1995) and are nowadays commonly used in the field of text classification and genre identification.[25] As stressed by Joachims (1998: 138), “[i]t is based on the *Structural Risk Minimization* principle from computational learning theory. The idea of structural risk minimization is to find a hypothesis *h* for which we can guarantee the lowest true error.”

In AGI, the principle at work is the following. Based on a set of training texts that are all marked as belonging to one specific genre, the SVM algorithm builds a model that will be later used in order to assign a genre to any new text (see 64). In its simplest linear form, this model is a representation of the texts as points in space; the texts that belong to one genre are mapped so as to be located as far as possible from texts belonging to another genre. In technical terms, the goal is to construct a hyperplane that separates the set of examples belonging to one category from the set of examples belonging to another category with the widest possible margin. Fig. 7 is an illustration of the basic SVM principle: the two groups (i.e. genre in AGI) of points (i.e. texts in AGI) are mapped respectively under and above the hyperplane; the hyperplane, which maximizes the margin (distance *a*) between

[25] SVM classifiers are known to be very accurate text classifiers, see e.g. Dumais *et al.* 1998; Joachims 1998; Dewdney *et al.* 2001; Basili & Moschitti 2005; Cleuziou & Poudat 2008.

the two groups, is represented by the unbroken line, while the points on the margin (b) are called support vectors.

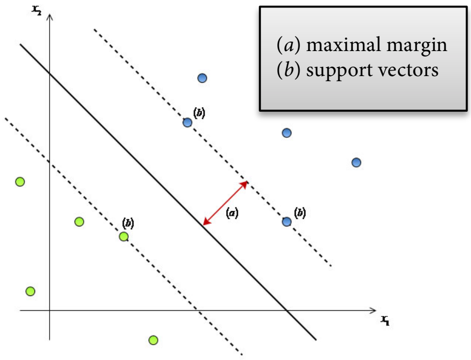


Figure 7. The principle of SVM

When the learning phase is completed, any new text is categorized in a genre depending on which side of the hyperplane it is mapped.

The software used to perform tests is called *SVM multi-class*. It allows, *inter alia*, classification to be performed with more than two classes as output (seven genres in the present case).[26] In the two case studies below, inputs are texts represented as vectors of lemmata and verbal inflexions respectively:[27] Each component of the vector corresponds to one lemma or verbal inflexion of the text and its value is weighted with the *tf-idf* (*term frequency-inverse document frequency)* function (see n. 16).

4.2.1. SVM with lemmata

Global performance for SVM classification based on lemma weighted with *tf-idf* is about 80.6%. This is — contrary to the expectations (see n. 25) — approximately 4% less than the results of the naïve Bayes classifier. As shown by the confusion matrix of Fig. 8, the performance of each genre is very close to the results obtained with the Bayes classifier, with the exception of the administrative texts (60%) being poorly recognized, which probably points to the fact that SVM needs a more extensive corpus in order to perform efficiently.

	LET.	JUD.	OR.	EDU.	MON.	HYM.	ADM.	**PERF.(%)**
LET.	133	5	0	0	1	4	0	**93.0**
JUD.	3	39	0	0	2	0	3	**83.0**
OR.	4	9	26	1	0	0	1	**63.4**
EDU.	2	0	0	25	1	0	1	**86.2**
MON.	0	0	0	0	28	0	0	**100.0**
HYM.	3	0	0	0	1	15	0	**78.9**
ADM.	3	2	0	0	1	0	9	**60.0**

Figure 8. Confusion matrix with SVM (lemmata as inputs)

[26] http://svmlight.joachims.org/svm_multiclass.html. A linear kernel has been used.

[27] One should stress here that the feature space has not been reduced for these two tests, which means that all the lemmata and verbal inflexions have been taken into account (for methods of feature selection, see Yang & Pedersen 1997). Furthermore, no removal of function words was performed.

4.2.1. SVM with verbal inflexion

Besides the semantic feature (based on the lemmata), we tested a morphological feature, taking as inputs the verbal inflexions found in the various genres. Although this criterion appears to be relevant in some cases (e.g. 84.6% of well classified letters), the performance with verbal inflexions as inputs is globally low, ca. 64%. Nevertheless, there is obviously a link between the types of verbal inflexions and genres, given that nearly two out of three texts are well classified with this criterion. Consequently, it appears that verbal inflexion could be used as a secondary criterion in combination with other more relevant criteria like the thematic one.

4.3. *The Segment and Combine method*

The last method of classification we tested for this first investigation of AGI in Late Egyptian is the so-called "Segment and Combine" method.[28] This is a generic method for supervised classification of *structured* objects. This means crucially that, unlike with the two classifiers described in §4.1 and §4.2, the syntactic organization of the texts is here taken into account.

The principle at work with this method is the following: (1) the texts are segmented in sequences of words, lemmata, inflexions, parts-of-speech, etc.; (2) a model (based on the training set) is learned that relates these sequences to a category (here a genre); (3) the texts (belonging to the test set) — that are considered as structured objects — are classified by combining the predictions made for their segments.

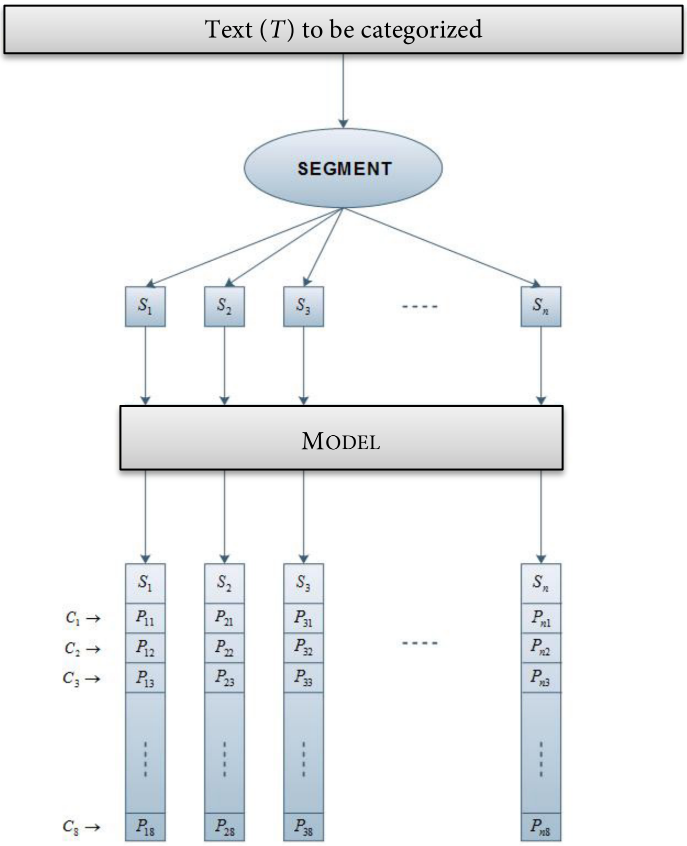


Figure 9. The Segment & Combine method (1)

The diagram of Fig. 9 illustrates the segmentation phase of the Segment and Combine method: (1) the text to be classified (T) is segmented in n sequences; (2) the n sequences are then submitted to the

[28] See Geurts & Wehenkel 2005; Geurts *et al.* 2005; 2006.

prediction function (i.e. the learned MODEL). The outcome is the assignment of a weighted vector to each sequence, where each vector component corresponds to one of the seven literary genres.

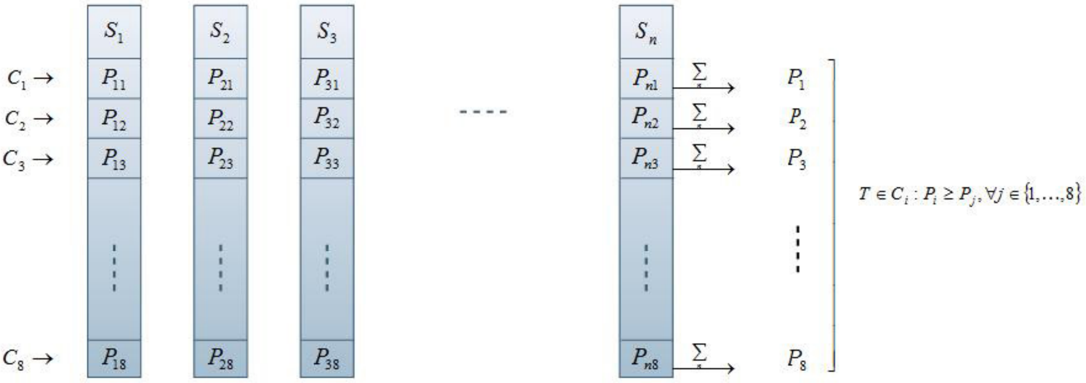


Figure 10. The Segment & Combine method (2)

The diagram of Fig. 10 illustrates the combination phase of the Segment and Combine method. For any text *T*, one proceeds with the addition of vectors (combination of the predictions made for each sequence); the result is a vector, each component of which corresponds to the weight associated with the respective literary genres. Finally, the text *T* is attributed to the genre that has the greatest weight.

The Segment and Combine method has been applied to the texts of the corpus on sequences made up of five words followed by a verb, itself followed by two words:[29] [w_1 w_2 w_3 w_4 w_5 VERB w_6 w_7]. The tests have been completed using *SVM multi-class*,[30] taking into account the lemmata (§4.3.1) and the parts-of-speech of $w_{1\text{-}7}$.

4.3.1. Sequence of lemmata

The performance of the method for sequences of eight lemmata (with a verb in the sixth position) amount to 67% of well classified texts. As shown by the confusion matrix in Fig. 11, however, the percentage varies significantly from one genre to another. Next to inaccurate results for some categories containing few and/or very short texts (oracular questions, hymns and prayers, and administrative texts), the Segment and Combine method gives excellent results for other categories (four genres have a performance higher than 90%). It should be stressed that this method performs better than the naïve Bayes classifier and the SVM method with the judicial documents (91,5%) and the educational texts (93,1%).

[29] Other tests with general sequences made up of 7, 9 and 11 words have been completed. These tests did not result in better performance. The performance increases slightly with the lemmata, but it decreases when considering the sequences of parts-of-speech.

[30] See n. 26.

	LET.	JUD.	OR.	EDU.	MON.	HYM.	ADM.	**PERF.(%)**
LET.	137	4	0	2	0	0	0	**95.8**
JUD.	2	43	0	0	1	0	1	**91.5**
OR.	25	4	11	0	0	0	1	**26.8**
EDU.	2	0	0	27	0	0	0	**93.1**
MON.	0	1	0	0	27	0	0	**96.4**
HYM.	4	2	0	2	1	10	0	**52.6**
ADM.	3	9	0	0	1	0	2	**13.3**

Figure 11. Confusion matrix for the Segment and Combine method (lemmata as inputs)

4.3.2. Sequence of parts-of-speech

The Segment and Combine method has also been applied to sequences of eight parts-of-speech (with a verb in the sixth position). The performance using this criterion is not good, with only 53.4% of texts correctly categorized. The size of the training sample might be at issue here. Indeed, the most populous genre (the letters) attracted the highest number of texts belonging to other genres. For example, none of the 29 educational texts are adequately classified, and 21 of them are predicted to be letters.

5. CONCLUSIONS

In this paper, we applied three supervised learning methods[31] in order to perform AGI within the Ramses corpus of Late Egyptian texts. The performance of each classifier is summarized in Fig. 12:

FEATURE	CLASSIFIER	PERFORMANCE
Lemmata	NBC	84,3%
	SVM	80,6%
Verbal inflexion	SVM	64%
Sequence of lemmata	S&C (SVM)	67%
Sequence of inflexions	S&C (SVM)	53,4%

Figure 12. Summary of the classification performances

The results are quite encouraging if one takes into account the fact that, contrary to most of the experiments in AGI:

(1) we only investigated isolated features (and not combinations of features);
(2) the size of the corpus on which the tests were performed is small;
(3) the number of categories (i.e. seven genres) is quite high.

Nevertheless, we observed that the genres that were sufficiently populated when the tests were completed (e.g. the letters) regularly exceed 90% of well classified texts (when the lemmata are used as features).[32]

Furthermore, the innovative Segment & Combine method is much promising: indeed, it outmatches the results of the naïve Bayes classifier and the SVM method with two genres when using sequences of lemmata.

[31] Naïve Bayes classifier (NBC); Support Vector Machine (SVM); Segment and Combine (S&C) method.

[32] This criterion apparently always works better than the *inflexion* (of verbs) and *part-of-speech* features.

Finally, the heuristic value of AGI in studying register variation in Late Egyptian carries out its function. The confusion underlined in §4.1.1, for instance, between the letters, the judicial and the administrative texts is evidently telling and points to a similarity between the registers that are actualized in these genres. Another case in point is the 100% correct categorization for the monumental royal texts (both with NBC and SVM), which shows the deep margin that separates the vocabulary of these texts from other written productions of the time.

To conclude, one cannot overemphasize the fact that the performances of the three classifiers tested in this paper could be considerably enhanced — for the purpose of developing efficient NLP tools like taggers or parsers — both by combining various linguistic features and by integrating extra-linguistic (i.e. material or structural, see §3) ones. As shown in §4.1.2, one could take into account the length of texts (as well as meta-data on documents, like the date of composition, the writing support, etc.), which would definitely improve the results of AGI in Late Egyptian.

BIBLIOGRAPHY

BASILI, Roberto & Alessandro MOSCHITTI. 2005. *Automatic Text Categorization from Information Retrieval to Support Vector Machine. A Text Book for Courses in Computer Science and Computational Linguistics*, Rome, University of Rome.

BEAUVISAGE, Thomas. 2001. Exploiter des données morphosyntaxiques pour l'étude statistique des genres — Application au roman policier, in: *Texto !*, available online at http://www.revue-texto.net/1996-2007/Inedits/Beauvisage/index.html.

BIBER, Douglas. 1986. Spoken and written textual dimensions in English: Resolving the contradictory findings, in: *Language* 62/2, p. 384-413.

—. 1988. *Variation across Speech and Writing*, Cambridge, Cambridge University Press.

—. 1993a. The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings, in: *Computers and the Humanities* 26/5-6, p. 331-345.

—. 1993b. Using register-diversified corpora for general language studies, in: *Computational Linguistics* 19/2, p. 243-258.

—. 1995. *Dimensions of Register Variation. A Cross-Linguistic Comparison*, Cambridge, Cambridge University Press.

CLEUZIOU, Guillaume & Céline POUDAT. 2008. Classification des textes en domaines et en genres en combinant morphosyntaxe et lexique, in: *Défi Fouille de textes (TALN '2008)*, available online at http://hal.archives-ouvertes.fr/hal-00466059.

DEBOLE, Franca & Fabrizio SEBASTIANI. 2004. Supervised term weighting for automated text categorization, in: Spiros SIRMAKESSIS (ed.), *Text Mining and its Applications. Results of the NEMIS Launch Conference* (= Studies in *Fuzziness* and *Soft Computing* 138), p. 81-98.

DEWDNEY, Nigel, Carol VANESS-DYKEMA & Richard MACMILLAN. 2001. The form is the substance: Classification of genres in text, in: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*.

DUMAIS, Susan, John PLATT, David HECKERMAN & Mehran SAHAMI. 1998. Inductive learning algorithm and representations for text categorization, in: *CIKM '98. Proceedings of the Seventh International Conference on Information and Knowledge Management*, ACM Press, p. 148-155.

GEURTS, Pierre & Louis WEHENKEL. 2005. Segment and combine approach for nonparametric time-series classification, in: *Lecture Notes in Computer Science* 3721 (= *Knowledge Discovery in Databases: Pkdd 2005*), Berlin, Springer Verlag, p. 478-485.

GEURTS, Pierre, Antia BLANCO CUESTA, Louis WEHENKEL. 2005. Segment and combine approach for biological sequence classification, in: *Proceedings IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2005)*, p. 194-201.

GEURTS, Pierre, Raphaël MARÉE & Louis WEHENKEL. 2006. Segment and combine: A generic approach for supervised learning of invariant classifiers from topologically structured data, in: *Proceedings of the Machine Learning Conference of Belgium and The Netherlands (Benelearn)*, p. 15-23.

JEBARI, Chaker. 2009. A new centroid-based approach for genre categorization of Web pages, in: *Journal for Language Technology and Computational Linguistics* 24/1, p. 73-96.

JOACHIMS, Thorsten. 1998. Text categorization with support vector machines: Learning with many relevant features, in: *Proceedings of 10th European Conference on Machine Learning*, p. 137-142.

KARLGREN, Jussi & Douglass CUTTING. 1994. Recognizing text genres with simple metrics using discriminant analysis, in: *Proceedings of the 15th International Conference on Computational Linguistics (COLING '94)*, Kyoto, p. 1071-1075.

KARLGREN, Jussi, Ivan BRETAN, Johan DEWE, Anders HALLBERG & Niklas WOLKERT. 1998. Iterative information retrieval using fast clustering and usage specific genres, in: *Proceedings of the Eighth DELOS Workshop on User Interfaces in Digital Libraries*, p. 85-92.

KESSLER, Brett, Geoffrey NUNBERG & Hinrich SCHÜTZE. 1997. Automatic detection of text genre, in: *Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, p. 32-38.

KO, Youngjoong, Jinwoo PARK & Jungyun SEO. 2002. Automatic text categorization using the importance of sentences, in: *Proceedings of the 19th International Conference on Computational Linguistics (COLING '02)*, vol. 1, p. 1-7.

—. 2004. Improving text categorization using the importance of sentences, in: *Information Processing and Management* 40/1, p. 65-79.

LIM, Chul Su, Kong Joo LEE & Gil Chang KIM. 2005. Multiple sets of features for automatic genre classification of web documents, in: *Information Processing and Management* 41, p. 1263-1276.

LEE, David Y.W. 2001. Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC Jungle, in: *Language Learning & Technology* 5/3, p. 37-72.

LEE, Yong-Bae & Sung Hyon MYAENG. 2002. Text genre classification with genre-revealing and subject-revealing features, in: *Proceedings of the 25th Annual International ACL-SIGIR Conference on Research and Development in Information Retrieval*, p. 145-150.

LEWIS, David D. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval, in: *Proceedings of ECML-98, 10th European Conference on Machine Learning*, Chemnitz, DE, p. 4-15.

MALRIEU, Denise & François RASTIER. 2002. Genres et variations morphosyntaxiques, in: *Texto!*, available online at http://www.revue-texto.net/Inedits/Malrieu_Rastier/Malrieu-Rastier_Genres.html.

MICHOS, Stefanos, Efstathios STAMATATOS, Nikos FAKOTAKIS & George KOKKONAKIS. 1996. An empirical text categorizing computational model based on stylistic aspects, in: *Proceedings of the Eighth International Conference on Tools with Artificial Intelligence*, p. 71-77.

POLIS, Stéphane, Anne-Claude HONNAY & Jean WINAND. Current volume. Building an annotated corpus of Late Egyptian. The Ramses Project: Review and perspectives.

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

SALTON, Gerard & Christopher BUCKLEY. 1988. Term weighting approaches in automatic text retrieval, in: *Information Processing and Management* 24/5, p. 513–523.

SANTINI, Marina. 2010. Cross-testing a genre classification model for the Web, in: Alexander MEHLER, Serge SHAROFF & Marina SANTINI (eds.), *Genres on the Web: Computational Models and Empirical Studies*, Dordrecht, Springer, p. 87-128.

SANTINI, Marina, Georg REHM, Serge SHAROFF and Alexander MEHLER. 2009. Automatic Genre Identification: Issues and Prospects, special issue of the *Journal for Language Technology and Computational Linguistics* 24/1.

SEBASTIANI, Fabrizio. 2002. Machine learning in automated text categorization, in: *ACM Computing Surveys* 34/1, p. 1-47.

STAMATATOS, Efstathios, Nikos FAKOTAKIS & George KOKKINAKIS. 2000a. Automatic text categorization in terms of genre and author, in: *Computational Linguistics* 26/4, p. 471-495.

—. 2000b. Text genre detection using common word frequencies, in: *Proceedings of the International Conference on Computational Linguistics (COLING 2000)*, p. 808-814.

VAPNIK, Vladimir. 1995. *The Nature of Statistical Learning Theory*, Dordrecht, Springer.

VIDULIN, Vedrana, Mitja LUŠTREK & Matjaž GAMS. 2009. Multi-label approaches to Web genre identification, in: *Journal for Language Technology and Computational Linguistics* 24/1, p. 97-114.

YANG, Yiming & Jan O. PEDERSEN. 1997. A comparative study on feature selection in text categorization, in: *ICML '97. Proceedings of the Fourteenth International Conference on Machine Learning*, San Francisco, Morgan Kaufamnn, p. 412-420.

Flexible Use of Text Annotations and Distance Learning[*]

Mark-Jan NEDERHOF
University of St Andrews

1. INTRODUCTION

A text may be analysed on various levels. If we restrict ourselves to types of analysis that are predominantly linear in nature, then we can distinguish for example:

- analysis of writing, orthography and palaeography,
- lexical analysis and morphology,
- syntactic analysis,
- semantic analysis.

The results of different kinds of analysis can be expressed in terms of appropriate text annotations. For example, for an analysis of the (hand-)written form of a document, the annotation may consist of a sequence of characters that express the interpretation of the physical appearance of the manuscript. Whereas for some writing systems and some kinds of documents this type of annotation may be straightforward, it is less so when the number of characters is very large, as in the case of Akkadian cuneiform or Chinese, and even potentially open-ended, as in the case of Ancient Egyptian, where the distinction between signs is not always clear-cut.

The problem is exacerbated by cursive styles of writing (cf. hieratic and demotic) or the poor conditions of manuscripts. An example of badly damaged manuscripts are some of the Herculaneum Papyri (Sider 2009). The term *transcription* is used for the representation of hieratic texts, usually found on papyrus, in terms of normalised hieroglyphs. Transcription is generally considered to be a form of interpretation, as a degree of uncertainty may be involved in identifying sign occurrences. A comprehensive overview of writing systems is offered by (Daniels & Bright 1996).

Related annotations include functional descriptions of hieroglyphs. In the case of Ancient Egyptian for example, one may want to distinguish between use of a hieroglyph as phonogram, as logogram, or as determinative. A sharp distinction between these three classes cannot be made, and some classes of hieroglyphs (e.g. phonetic determinatives) do not fit well in any of these classes. In this regard, the introduction to the sign list on pp. 438–441 of Gardiner (^{3}1957) is enlightening.[1]

Nevertheless, with the understanding that a small number of occurrences of hieroglyphs may be hard to classify, one may systematically annotate texts by indicating the function of each hieroglyph.

[*] I gratefully acknowledge discussions with Serge Rosmorduc about automatic processing of hieroglyphic texts. Many thanks go to anonymous reviewers for a large number of improvements to the text. The tool used for joint translation of the wisdom text of Ptahhotep contains PHP code contributed by Geoffrey Watson. The tool described in this paper was partly developed thanks to the generous assistance of a fellowship from the Leverhulme Trust.

[1] The problems become worse if one considers finer distinctions between functions of hieroglyphs, for example following Schenkel (1971); see also Schenkel 1984.

Statistical analysis of such annotations may subsequently reveal insights about the writing system. For an example of such statistical analysis, see the introduction of Hannig (1995: XXXIV-XXXV).

Words in a text may be annotated by their morphological structure, their grammatical function, and annotation with lemmas may link word occurrences to a lexicon. A particular kind of lexical annotation that is very useful to language learners is a gloss, a literal translation for each word individually.

Syntactic analyses may for example take the form of parse trees or dependency structures. Such linguistic annotations have become commonplace in linguistic research involving modern languages (Jurafsky & Martin 2000), but use in the field of philology is relatively infrequent. A notable exception is the Ramses project involving Ancient Egyptian, which is discussed elsewhere in this volume.

Sentence annotations may comprise logical formulas, or other kinds of semantic or pragmatic information. For philological purposes, a very useful representation of the meaning of a portion of text is simply a translation in a modern language. Such a translation may be quite literal in order to clarify the grammatical structure of the original text, or it may be more free in order to clarify the interpretation of a portion of text in its context.[2]

Whereas different levels of annotation may exist independently, the information they carry can be intertwined across levels. For example, two different interpretations of an occurrence of a hieroglyph can lead to two widely different syntactic analyses and semantic interpretations of a portion of text. Conversely, aspects of an interpretation of a text on a higher level of annotation may justify annotations on a lower level. Therefore, it may enhance complete understanding of a text if different levels of annotation can be easily studied and compared, one next to the other.

The combination of several linear forms of annotation in one unified representation is called *interlinear text* (see Bow *et al.* 2003). The individual annotations within interlinear text are called *tiers*. Typically, the text is divided into paragraphs, sentences or phrases, and for each, the corresponding parts of the respective tiers are printed closely together. The exact arrangement can be one tier under the other, or one tier next to the other. If tiers are printed beneath each other, the horizontal placement of elements may be chosen so as to align corresponding elements in the different tiers. This helps the scholar to understand the relation between the different levels of annotation.

Interlinear text is widely used, for example for teaching modern languages, and for documenting endangered languages. Some tools allow combination with audio and video material (Wittenburg *et al.* 2006). Despite the sophistication of many viewing tools for multi-tier text annotation, there is often an implicit assumption that all annotations can be anchored on an unchanging representation of a text, or even that all levels of annotations are integrated into a single file created by one linguist or a small team of linguists.[3]

The benefit of this assumption is clear: the representation of a closely-linked collection of tiers allows a relatively straightforward grouping of corresponding elements from each tier. Subsequently, interlinear text may be created by printing these corresponding elements closely together in columns or rows, phrase by phrase, or sentence by sentence, from the beginning to the end of the text.

In the domain of philology however, a number of specific obstacles arise. First, a single representation of a text on which all annotations can be anchored is often difficult to obtain. For example, one may be tempted to store a sentence of translation of a hieroglyphic text together with an indication that the sentence covers the i-th hieroglyph up to the j-th hieroglyph. However, in the case of damaged text, scholars may disagree how many signs can still be clearly discerned, or how many damaged or entirely lost signs can be reconstructed with certainty in the light of the context. What is the i-th hieroglyph by one interpretation may be the $(i + 1)$-th by another.

[2] See for example Munday (2001) for different approaches to translation.

[3] See also Bird & Liberman 2001.

Second, there is often considerable uncertainty about the correct interpretation of ancient texts. In the case of Ancient Egyptian, scholars may even disagree how to segment a sequence of hieroglyphs into words. In such a case, arbitrarily including only one interpretation in interlinear text and excluding all dissenting views may not be beneficial to the free exchange of ideas.

Third, for a considerable number of texts there are several text variants, some from different periods. The inclusion of several text variants in one interlinear text encourages a deeper understanding of both the text and diachronic linguistic processes. In addition, a single representation of a text on which other representations can be anchored is difficult to obtain, especially if none of the extant manuscripts cover the entire text. A classical example is the text of "The Eloquent Peasant" (Parkinson 1991).

The above observations suggest that electronic resources encoding levels of annotation of ancient texts should follow principles different from those we would use for modern texts. Rather than having several annotations of one text closely linked to one another or unified in a single data format, a more distributed approach seems in order. That is, scholars produce separate electronic resources, possibly for different interpretations of a text, possibly dealing with different text variants, without having to agree with one another on how to segment the text into sentences, phrases, words or orthographic units.

Given these considerations, it is not altogether simple to build software capable of visualising the ensemble of available resources for a given text. For example, suppose we have two tiers consisting of different translations. Assuming the translations were produced independently by different scholars, then interlinear text cannot be readily created. First it needs to be established which sentences from the two translations belong together. In cases where the segmentation of the text into sentences is different for the two translations, a correspondence between the two tiers may be found in terms of smaller linguistic units, e.g. phrases. However, differences in word order between two translations of one ancient text may preclude a fine-grained correspondence between small linguistic units in general. Differing word order is particularly a problem when the two translations of one ancient text are in different modern languages, say one in English and the other in German.

Finding correspondences between two linear structures is called *alignment*. Alignment can take the form of n-to-m mappings, for example indicating that n sentences in one tier correspond to m sentences in a second tier. In the simplest case $n = m = 1$, but in practice one would at the very least also need 2-to-1 and 1-to-2 mappings in order to deal with the case of two translations having different numbers of sentences. An alternative to n-to-m mappings is to indicate positions in two tiers that correspond. For example, one may require that the first word of a sentence in the first translation and the first word of a sentence in the second translation must be printed one below the other. It is this kind of alignment that we will discuss in the present article.

Alignment may be manual or automatic. In the case of Ancient Egyptian, automatic alignment is particularly effective for hieroglyphic encoding and transliteration in the Egyptological transliteration alphabet. Initial experiments reported by Nederhof (2008) suggest that reliable alignment can be obtained on the level of individual words, using very simple models of hieroglyphic writing. Alignment of hieroglyphic encodings of text variants is also relatively straightforward, assuming variation between texts is not too great. More difficult is the automatic alignment of different translations. This problem has received considerable attention in computational linguistics.[4]

Manual alignment can replace or complement automatic alignment, for example when two textual resources have been created by two scholars, and a third scholar explicitly links the tiers in the resources together. When a resource consisting of several tiers is created by a scholar, then the tiers may be manually aligned as a consequence of the file format. For example, the text may be segmented into sentences or phrases, and for each such unit, the file contains the corresponding elements from

[4] See Och & Ney 2003.

the respective tiers. This in effect links the tiers together. Examples will be provided in following sections.

2. FLEXIBLE USE AND REUSE OF LINGUISTIC RESOURCES

In light of the above motivations, we can consider a scenario whereby an annotated corpus is created, as follows:

- A minimal set of requirements of a file format are assumed. The file format is simple enough to convert other formats into, and existing printed documents can be digitised in this form without reinterpretation. That is, digitisation can to a large extent be done by technical assistants rather than by scholars.
- Two scholars creating two files annotating the same text do not need to agree on common conventions or common interpretations, for example how to segment the text or what transliteration alphabet to use.
- The software to visualise the available resources is sophisticated, includes automatic alignment, and allows flexible rendering based on various preferences: which fonts to use, which tiers to show, whether to render on the screen or on paper, etc.
- Once created, the electronic resources can be reused, without requiring further manual manipulation.

This approach should be contrasted with a more traditional approach of creating annotated corpora, which can be described as follows:

- Funds are secured.
- A team of scholars is formed and employed for a number of years.
- Agreements are made about the scope of texts to be included, the level of annotation, the annotation conventions, the handling of contentious cases, etc.
- During the development of the corpus, techniques of quality assurance are implemented to guarantee high accuracy and consistency.
- After the work is completed, the corpus is made public, and the team is disbanded.

Examples of modern corpora constructed along these lines are the Penn Treebank (Marcus *et al.* 1993) and the British National Corpus (Leech *et al.* 1994), both of English, and the Negra Corpus (Skut *et al.* 1997) of German. Several corpora also exist for ancient languages, such as the Perseus corpus (Crane 1998; Crane & Rydberg-Cox 200) of Ancient Greek and Latin and the ETCSL corpus (Ebeling & Cunningham 2007) of Sumerian. The Ramses corpus of Late Egyptian is discussed in Polis, Honnay & Winand (current volume).

This approach is by far the most desirable if the objective is to obtain a corpus that is highly accurate, consistent, and systematically covers a predetermined set of texts. If it is imperative that all these desiderata be fulfilled, then there may in fact be no real alternative.

However, this traditional method for the creation of corpora also has many disadvantages. First, the labour costs are very high. For well established areas of philology, there will be unavoidable duplication of effort, in that many types of annotation, and especially translations, are already available in printed form.

In addition, it is often difficult to maintain a corpus after the team who developed it has been disbanded. Maintenance may include correcting mistakes that were found, or it may involve adding new texts or new levels of annotations of existing texts, which may be hard to incorporate with design decisions made by the original project.

Lastly, where there are competing ‘schools’ of grammar, the team developing the corpus may receive criticism of being biased towards one school, ignoring dissenting analyses.

The less centralistic approach that we propose for development of electronic resources may offer at least partial solutions to these problems. It does not require a single costly project, nor scholars exclusively dedicated to text annotation for a considerable period, because the work can be divided over the entire community. Annotation can be flexible, allowing for appropriate levels to be created for the relevant texts.

In addition, printed annotations can be digitised to become accessible to large numbers of scholars. In areas such as Ancient Egyptian and Assyrian philology, students and scholars who are not affiliated with institutions with the necessary libraries are often confronted with great difficulty getting hold of relevant publications. Because of the present economic climate, fewer and fewer centres of study offer courses in ancient languages, and thereby it also becomes increasingly difficult for students to consult experts who are able to help them with learning ancient languages. It is likely therefore that there will be a growing need for electronic textual resources that can be accessed over the Internet.

One disadvantage of the decentralised approach to the creation of corpora is that it is difficult to ensure a uniform degree of consistency and accuracy. For applications such as large-scale lexicography and statistical linguistic research, these may be prohibitive obstacles. However, for many applications that pertain to individual texts, most users may benefit from any available electronic resources. As long as the provenance and reliability are made clear, it is less important that all resources have the same reliability, are drawn from the same sources, or follow the same notational conventions.

A more significant disadvantage is that sophisticated software is needed to process text annotations coming from various sources, and render them in a uniform interlinear format, so that users can study a text in a convenient manner. It is the objective of this article to show that the required software can be realised, and our proof-of-concept is a discussion of data formats together with the introduction of a working tool.

Our domain will be Ancient Egyptian hieroglyphic texts. This domain is particularly pertinent for the problems considered in this article because of the use of transliteration, which often forms an additional tier of annotation between an encoding of the manuscript and its translation. Also of special interest are the Ancient Egyptian texts that have survived in several variants. Cases where segmentation of a sequence of hieroglyphs into words is uncertain pose a further challenge to the creation of interlinear text incorporating different interpretations.

3. THE SOFTWARE

The current implementation of the software refines earlier designs, the first of which was reported by Nederhof (2002a). The present implementation[5] language is the programming language Java, which runs on all major platforms such as Windows, Mac OS X and Linux.

Java allows an objected-oriented software design. Among the advantages this offers, of particular relevance here is the ability to describe data structures and algorithms in an abstract manner, omitting details that can be filled in later, or that can be filled in in several different ways.

Concretely, the largest portion of the program code deals with concepts such as textual resources consisting of several tiers, interlinear text, constraints on the formatting of the tiers, and algorithms to solve those constraints to result in suitable interlinear text, as has been outlined previously in Nederhof (2009). The user has a choice which tiers from the available resources are to be displayed as part of the interlinear text, and there is an option to print the interlinear text to a PDF file.

There is also a simple infrastructure to maintain indexes of texts, and to import and export language resources for texts. None of this code however refers to any particular language (e.g. Ancient Egyptian or Akkadian) nor to any particular writing system (e.g. hieroglyphs or cuneiform). We will call this part of the program ‘Philolog’.

[5] The program can be downloaded from: http://www.cs.st-andrews.ac.uk/~mjn/egyptian/align/.

A smaller portion of the program code specifies the language and writing system of Ancient Egyptian. This includes code and fonts to edit, render and analyse hieroglyphic text as well as transliterations in the Egyptological transliteration alphabet. The complete software package is called ‘PhilologEg’, which can be seen as an instantiation of ‘Philolog’. As a consequence of this design, it is straightforward to create other instantiations for other languages and writing systems, by replacing a relatively small portion of code.

In addition, the modular design makes it easy to add new kinds of annotation. For example, it would be easy to add syntactic annotation to the tool without changing any of the existing design.

The tool manipulates textual resources. These resources may exist as files on the local file system, and referred to as path names. These resources may then be read as well as modified. However, the resources may also exist as web addresses (URLs). This allows for the possibility of editing one’s own translation on the local file system, visualised in interlinear text underneath a hieroglyphic encoding that is downloaded from the internet.

The type of hieroglyphic encoding implemented in PhilologEg is the Revised Encoding Scheme (RES). In Nederhof (2002b, 2008 and current volume) we have outlined arguments in favour of RES, as opposed to the most widely used encoding known as the Manuel de Codage. PhilologEg includes a graphical editor for RES, which allows hieroglyphic encoding to be visualised and manipulated in terms of tree structures, as illustrated in Fig. 1.

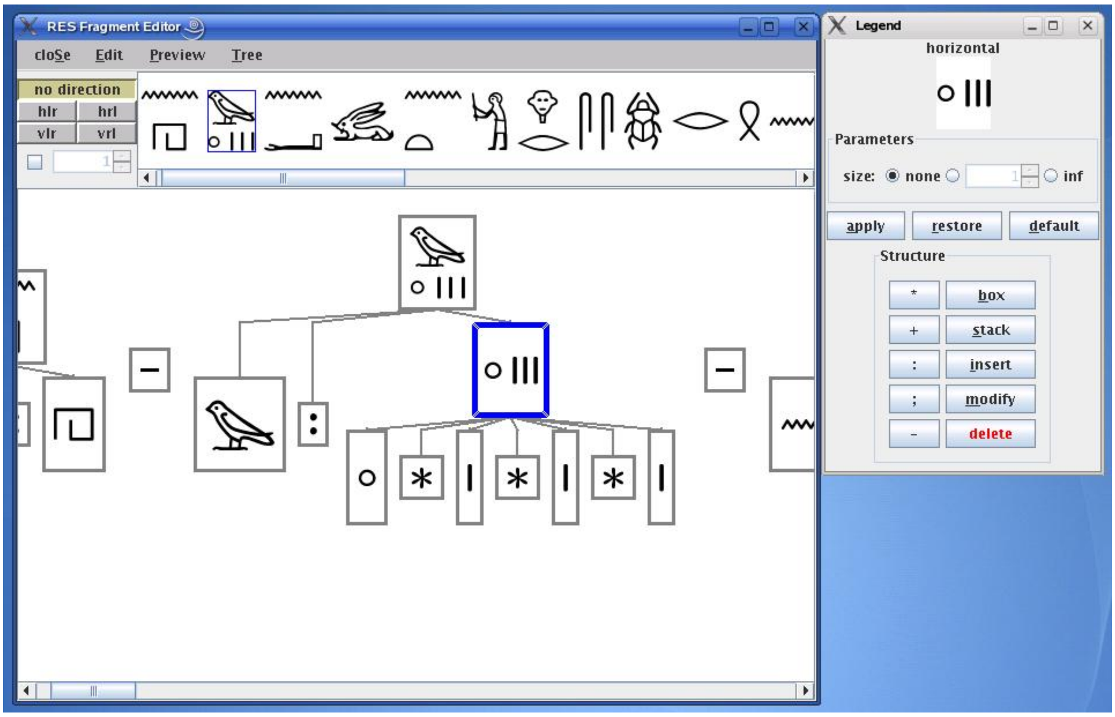


Figure 1. Graphical editor for RES[6]

4. THE DATA

This section discusses the main features of the data formats. We abstain from a complete listing of constructions, which does not seem desirable given that minor modifications may still be made in the near future in response to new insights.

[6] The bottom panel shows expressions as hierarchical structures. Each node in a structure shows the appearance of a subexpression. The panel above that shows the appearance of a complete fragment of hieroglyphic. The panel appended on the right allows editing of parameters of the hieroglyph or operator that is the current focus, as well as structural modifications taking place at the focus, such as insertion of new nodes.

All the data can be represented in the form of XML files (Harold & Means ^{3}2004). However, a so-called 'light' data format can also be used as an alternative. Its purpose is to simplify manual input of text without resorting to graphical editors. We will not further discuss the light data formats in this article.

The XML files represent information content rather than formatting. For example, no representation exists in the data format for an explicit line break. It is the task of the visualisation software to determine where line breaks should occur, subject to various constraints.

4.1. *Individual resources*

In the simplest case, there is a single electronic resource written by a single author. This is represented as a file containing one or more levels of annotation for a single text. These levels can be any combination of:

- hieroglyphic encoding,
- transliteration,
- translation,
- lexical annotation, comprising glosses, lemmas in dictionaries, parts of speech, etc.

For convenience, the text can be divided into segments, and the author may edit one segment at a time. A segment may be a phrase in the linguistic sense, but it can also be any unit of text that is convenient for the user to edit. The body of a file consists of zero or more such segments.

An example of a resource containing only hieroglyphic encoding is given in Fig. 2, together with a possible rendering. The exact rendering may depend on various parameters, such as the width of a window or the width of a printed page. In particular, line breaks appear when this width is exhausted, which is not necessarily at the end of a segment, nor at the end of a line in the input file. It should further be noted that the hieroglyphic encoding indicates by the construction '`[hrl]`' that the manuscript is horizontal right-to-left, but the tool changes this to left-to-right, to accommodate for alignment with other resources, as illustrated below.

```
<segment>
<hieroglyphic>
<coord id="2.6"/>[hrl].:a-wn:n-
  t:xt-Z1*Z1*Z1:n-.:t-
  tA:N23*Z1-E1:.*Z1*Z1*Z1*.-
  N25
</hieroglyphic>
</segment>

<segment>
<hieroglyphic>
<coord id="2.7"/>[hrl]F26-n:nw-
  w:t-F27-Z1[shade]*[shade]Z1*Z1:n-
  .:t-b-bA-[sep=0.3]A-[sep=0.2]
  Z7[e]-[shade]F27:Z1*Z1*Z1
</hieroglyphic>
</segment>
```

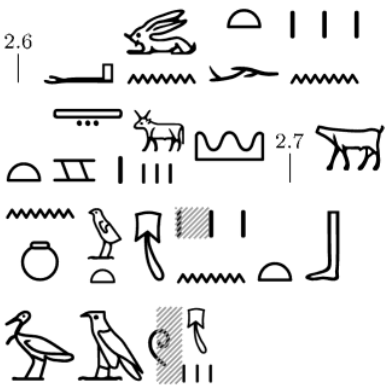


Figure 2. Part of the body of an electronic resource containing only hieroglyphic encoding, and a typical rendering by the tool

An example of a resource containing transliteration and translation is given in Fig. 3. As before, line breaks in the rendered interlinear text are not determined by line breaks within the input files, but by various constraints on the rendering process, such as the available width. In addition, there is horizontal alignment between the two tiers, induced by two sources of information. First, for each segment, there is alignment of the first elements of both tiers in that segment, as for example `jr` and 'Beware'. Second, there is alignment for coordinates, such as physical line numbers in the manuscript, for example '8.5'.

```
<segment>
<alphabetic>
Dd.jn ^nmtj-nxt pn
</alphabetic>
<translation>
This Nemtinakht then said:
</translation>
</segment>

<segment>
<alphabetic>
jr hrw <coord id="8.5"/> sxtj
</alphabetic>
<translation>
Beware, <coord id="8.5"/> peasant,
</translation>
</segment>
```

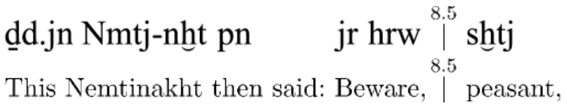

ḏd.jn Nmtj-nḫt pn jr hrw 8.5 | sḫtj
This Nemtinakht then said: Beware, 8.5 | peasant,

Figure 3. Part of the body of an electronic resource containing transliteration and translations, and typical rendering by the tool

Alignment may also be indicated manually, by so-called precedence constraints. One such constraint says that one position in one tier must come before (or at the same horizontal position as) a second position in a second tier. An example is shown in Fig. 4. There are symbolic labels for positions in the two tiers, and for example `<prec id1="26" id2="29"/>` indicates that position '26' must come before position '29'. Together with the converse `<prec id1="29" id2="26"/>` this means that the two positions must be aligned one under the other. This example also shows the use of footnotes; appropriate (unique) footnote markers are determined by the rendering tool.

```
<segment>
<alphabetic>
jbsA <pos id="26"/>jnbj
  <pos id="27"/>mnw
</alphabetic>
<translation>
wild mint,<note>Meaning
  is uncertain.</note>
<pos id="29"/>hedge plants,
<pos id="30"/>pigeons,
</translation>
<prec id1="26" id2="29"/>
<prec id1="29" id2="26"/>
<prec id1="27" id2="30"/>
<prec id1="30" id2="27"/>
</segment>
```

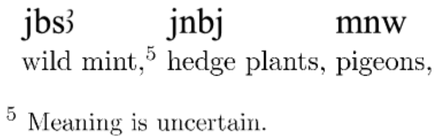


Figure 4. Manual alignment using precedence constraints

A resource may further contain:

- name of the author,
- date of creation and date of last change,
- a free-text description of what the resource represents, how it was obtained, which annotation conventions were used, etc.,
- optionally, the name of the text variant and the numbering scheme (see §4.4)
- a list of bibliographic references,
- optionally, information used for automatic uploading (see §5).

4.2. *Automatic alignment of resources*

We now consider a more complicated case, namely that we have two resources, which are two files created independently. The first file may for example contain hieroglyphic encoding, and the second file may contain transliterations and translations for the same text. The respective authors of the two resources may have included coordinates referring to physical line numbers in the manuscript. These may help in correctly aligning the first tier of hieroglyphic with the second tier of transliteration.

However, coordinates may be absent or may be too few to ensure that line breaks for the three tiers occur in corresponding positions. For this reason, the software includes automatic alignment of hieroglyphic and transliteration, which looks at the possible functions and meanings of hieroglyphs and relates them to sequences of letters in the transliteration alphabet. The tool then places corresponding line breaks for the two relevant tiers.

The outcome is shown in Fig. 5. Alignment of the hieroglyphs with the transliteration is done according to the coordinates, as for example '74' and '75' in the figure. In addition, a line break within the hieroglyphs is inserted to correspond with the segmentation of the translation and transliteration, so that the corresponding elements from the three tiers occur closely together.

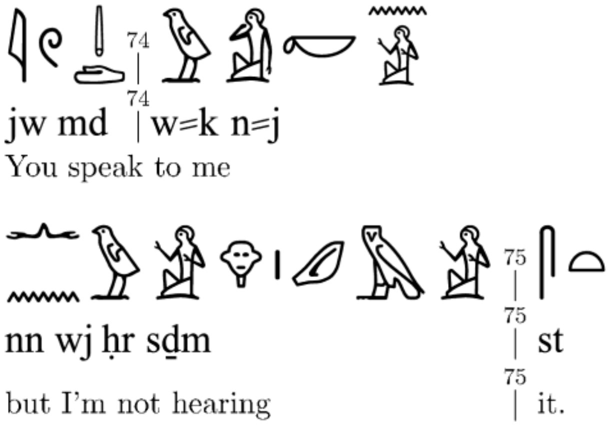


Figure 5. Interlinear text obtained from two independently created resources

Similarly, we have implemented a simple form of automatic alignment of different translations, for example, for two independently created resources containing translations in English and Dutch, respectively. The tool will try to break lines in corresponding positions. The automatic alignment is based on a heuristic that assumes that the number of words in segments of the first translation is comparable to that in corresponding segments of the second translation. More sophisticated forms of alignment are possible however.

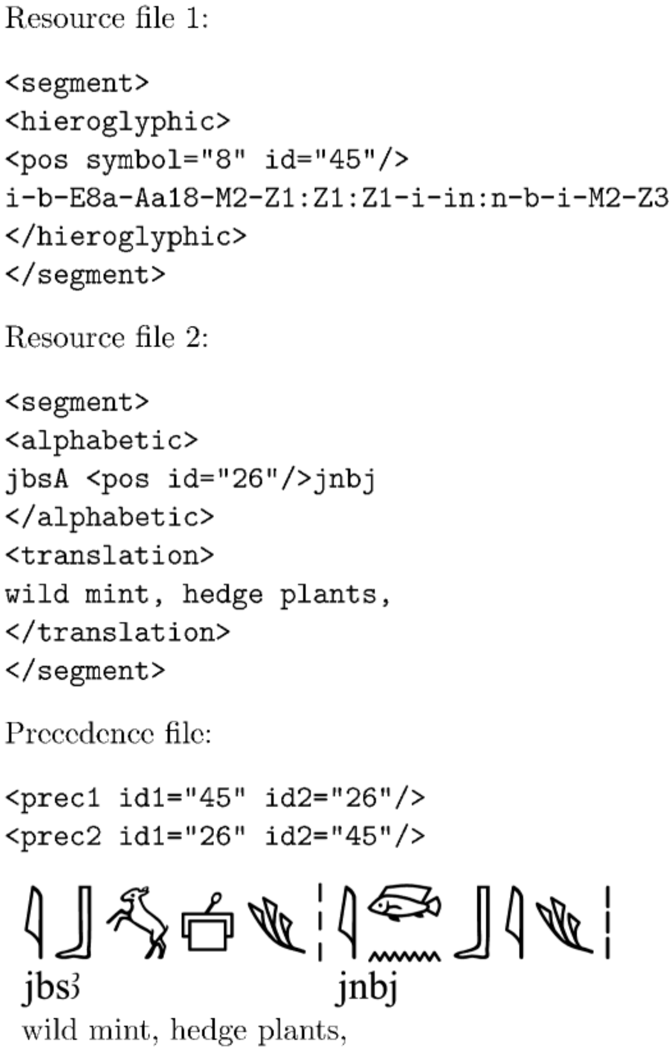
The modular design of the software allows new alignment algorithms to be added without changing the remainder of the program code. For example, if one were given Java code to do automatic alignment of French and English, this could be 'plugged in' into the existing tool to align

French and English translations. The tool falls back on simple heuristics if specialised alignment algorithms for certain types of tiers are not available.

Note that any alignment that is automatic may make mistakes, but the worst consequence of a mistake is that the interlinear text contains confusing line breaks. The rendered form is never factually incorrect.

4.3. *Manual alignment of resources*

Where automatic alignment is not available, or is not precise enough to guarantee a high accuracy, a user may also connect two tiers from different resources by manually indicating precedence constraints between positions in two tiers. Manual precedence constraints always override automatic alignment. In the general case, the two original resources may be read-only, and may even be accessed as web addresses on different sites. For this reason, the precedence constraints linking two resources are stored in a third file, possibly on the local machine of the user. Assuming that suitable symbolic names for the relevant positions already exist in the two resources, this idea is illustrated in Fig. 6. The element `<pos symbol="8" id="45"/>` gives a symbolic name '45' to the hieroglyph with index '8' in the following encoding. The precedence file links this to position '26' in the second resource file. The tag names 'prec1' and 'prec2' indicate two different directions of precedence constraints between the two resources. As these constraints on '45' and '26' exist in both directions, `jnbj` is aligned directly underneath the corresponding hieroglyphs.



Resource file 1:

```
<segment>
<hieroglyphic>
<pos symbol="8" id="45"/>
i-b-E8a-Aa18-M2-Z1:Z1:Z1-i-in:n-b-i-M2-Z3
</hieroglyphic>
</segment>
```

Resource file 2:

```
<segment>
<alphabetic>
jbsA <pos id="26"/>jnbj
</alphabetic>
<translation>
wild mint, hedge plants,
</translation>
</segment>
```

Precedence file:

```
<prec1 id1="45" id2="26"/>
<prec2 id1="26" id2="45"/>
```

Figure 6. Manual alignment between two resources

The reason for the use of symbolic names rather than absolute positions in the tiers is that we would like the precedence constraints to keep their validity when the original resources undergo (minor) changes by their authors. If an author removes a symbolic name altogether, then the worst that can happen is that a precedence constraint becomes without meaning and will be ignored.

What happens when one tries to put precedence constraints on two positions that are not associated with symbolic names depends on whether the resources can be edited. If they can be edited, then the tool automatically inserts a 'pos' tag with a new and unique symbolic name. If the resources cannot be edited, then the precedence constraints refers to a symbolic name nearest to the relevant position with an indication that a certain number of positions should be added or subtracted. For example, `<prec1 id1="A" id2="B" plus1="5" plus2="-10"/>` indicates that the position that occurs 5 symbols after symbolic position 'A' in the first resource file should be placed to the left of the position that occurs 10 symbols before symbolic position 'B' in the second resource file.

4.4. *Text variant and number scheme*

We assume that all tiers within one resource refer to the same text variant. The name of the text variant can be indicated in the file. In addition, it may be required to indicate which 'numbering scheme' the coordinates in a resource refer to.

To demonstrate this, we consider the text of The Eloquent Peasant. It has survived in four manuscripts. In manuscript R, the lines used to be numbered 1 to 229, but later publications use line numbers 1.1 to 31.8. To ensure correct alignment of two resources using different number schemes, we have added a file format with the sole purpose of equating line numbers in different schemes. It may contain lines such as `<map first="229" second="31.8"/>`.

5. LEARNING AND TEACHING

If linguistic data is represented in well-chosen file formats, then this data can be used and reused in a flexible manner for many different purposes. In this section we address possible use of our data formats and software for learning and teaching. Of particular interest is distance learning, which, as we argued in §2, is becoming increasingly important, as fewer and fewer institutions offer conventional classroom courses in ancient languages.

As we have explained before, the software we have developed allows the selection of tiers to be rendered. For the preparation of teaching material, teachers may choose to omit the tiers (most frequently transliteration and translation) that they want students to fill in as an exercise. In the typical case, only the tier of hieroglyphic text will then be printed. When the teaching material is to be printed on paper, an adequate amount of white space can be left for the students to fill in their interpretations. If teachers want to give hints to the students how to segment a text into phrases, these hints can be automatically produced out of the phrases of an existing translation for the text at hand.

Electronic teaching material can be prepared in much the same way, but with the possibility that the students use a graphical editor to add their transliterations and translations below an existing hieroglyphic encoding. Furthermore, an additional tier can be created in which a tutor puts comments, as feedback on translations submitted as coursework.

We will now discuss one test case of 'computer-aided collaborative learning' that we have developed recently. Its purpose is to help in the joint translation of the wisdom text of Ptahhotep (Žába 1956). This joint project is done via the Ancient Egyptian Language email list (Wilson 1997).

Conventional joint translations done with the help of email lists suffer from the following problems:

- Submitted translations exist as separate email messages in the mail boxes of subscribers. This makes it hard to keep track of which parts of a text have been translated already, by whom, and what the differences between the various interpretations are.

- There are technological difficulties using e.g. hieroglyphic writing within emails, and submitted translations cannot readily be compared to the original hieroglyphic text itself.

In order to address these problems we have created a tool that can be used by each subscriber to create and edit their own interpretation, using a graphical editor that places transliterations and translations immediately below given hieroglyphic text. When a user is ready to share an interpretation, they can upload it onto a central server, where it is combined with interpretations by others. Interlinear text is automatically created that shows one tier of hieroglyphic followed by pairs of tiers of transliteration and translation, one pair for each subscriber. Footnotes can be added to clarify interpretations. The technical realisation is outlined as follows:

- A so-called JAR file has been made available on a central web page. This file represents Java code packed together with all needed data, including hieroglyphic fonts and the hieroglyphic encoding of the text to be translated.
- The application has been developed to run on all major platforms, and in particular Windows, Mac OS X and Linux, and has been thoroughly tested on all of these platforms.
- Activation normally proceeds by a simple mouse click on the JAR file.
- The first time the tool is activated, the user is asked for their name, email address and a password distributed via the email list. The password serves to prevent abuse of the tool. The name and email address serve to distinguish subscribers. A local copy of a file is also created that will contain the interpretation of the text.
- After a new part of the interpretation has been entered in the graphical editor, a user presses an ‘upload’ button, which automatically sends the interpretation to the central server.
- At the central server, a PHP script verifies the password and stores the interpretation.
- A Java applet on the server allows all stored interpretations to be accessed and rendered as interlinear text, as explained earlier.
- For users who cannot use or do not want to use applets, the interlinear text is also made available as PDF file.

In this particular instance, the text is already segmented into ‘verses’, following the verse numbers of Žába (1956). However, for texts where such a segmentation is not available, the software allows students to segment the hieroglyphic text in whatever way they choose and attach translations to the chosen segments.

6. Conclusions

We have investigated the possibility of creating a corpus of text annotations through distributed efforts. We have presented software that is able to combine and visualise available textual resources in a meaningful way. This enables flexible use and reuse of textual material and enhances the possibilities for the study of texts. The approach is of particular interest to areas of philology where there are large numbers of text, but relatively few electronic resources readily available, such as in the case of Ancient Egyptian.

Bibliography

BIRD, Steven & Mark LIBERMAN. 2001. A formal framework for linguistic annotation, in: *Speech Communication* 33, p 23-60.

BOW, Cathy, Baden HUGHES & Steven BIRD. 2003. Towards a general model of interlinear text, in: *Workshop on Digitizing and Annotating Texts and Field Recordings*, Michigan State University.

CRANE, Gregory. 1998. The Perseus project and beyond: How building a digital library challenges the humanities and technology, in: *D-Lib Magazine* 4(1), p. 1-18.

CRANE, Gregory & Jeffrey A. RYDBERG-COX. 2000. New technology and new roles: The need for 'corpus editors', in: *Proceedings of the fifth ACM conference on Digital libraries*, San Antonio, Texas, United States, p. 252-253.

DANIELS, Peter T. & William BRIGHT (eds.). 1996. *The World's Writing Systems*, New York, Oxford University Press.

EBELING, Jarle & Graham CUNNINGHAM (eds.). 2007. *Analysing Literary Sumerian: Corpus-based Approaches*, Equinox Publishing.

GARDINER, Alan H. ^{3}1957. *Egyptian Grammar: Being an Introduction to the Study of Hieroglyphs*, Oxford, Griffith Institute.

HANNIG, Rainer. 1995. *Großes Handwörterbuch Ägyptisch-Deutsch: die Sprache der Pharaonen (2800-950 v.Chr.)*, Mainz, Philipp von Zabern.

HAROLD, Elliotte Rusty & W. Scott MEANS. ^{3}2004. *XML in a Nutshell*. 3rd ed., O'Reilly.

JURAFSKY, Daniel & James H. MARTIN. 2002. *Speech and Language Processing*, Prentice-Hall.

LEECH, Geoffrey, Roger GARSIDE & Michael BRYANT. 1994. CLAWS4: The tagging of the British National Corpus, in: *The 15th International Conference on Computational Linguistics*, vol. 1, p. 622-628.

MARCUS, Mitchell P., Beatrice SANTORINI & Mary Ann MARCINKIEWICZ. 1993. Building a large annotated corpus of English: The Penn Treebank, in: *Computational Linguistics* 19(2), p. 313-330.

MUNDAY, Jeremy. 2001. *Introducing Translation Studies: Theories and Applications*, London & New York, Routledge.

NEDERHOF, Mark-Jan. 2002a. Alignment of resources on Egyptian texts based on XML, in: *Proceedings of the 14th Table Ronde Informatique et Égyptologie*. On CD-ROM.

—. 2002b. A revised encoding scheme for hieroglyphic, in: *Proceedings of the 14th Table Ronde Informatique et Égyptologie*. On CD-ROM.

—. 2008. Automatic alignment of hieroglyphs and transliteration, 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*, Gorgias Press, p. 71-92.

—. 2009. Automatic creation of interlinear text for philological purposes, in: *Traitement Automatique des Langues* 50(2), p. 237-255.

OCH, Franz Josef & Hermann NEY. 2003. A systematic comparison of various statistical alignment models, in: *Computational Linguistics* 29(1), p. 19-51.

PARKINSON, Richard B. 1991. *The Tale of the Eloquent Peasant*, Oxford, Griffith Institute.

POLIS, Stéphane, Anne-Claude HONNAY & Jean WINAND. Current volume. Building an annotated corpus of Late Egyptian. The Ramses Project: Review and Perspectives.

SCHENKEL, Wolfgang. 1971. Zur Struktur der Hieroglyphenschrift, in: *Mitteilungen des deutschen archäologischen Instituts, Abteilung Kairo* 27, p. 85-98.

—. 1984. Schrift, in: Wolfgang HELCK & Wolfhart WESTENDORF (eds.), *Lexikon der Ägyptologie*, Wiesbaden, Harrassowitz, vol. 5, p. 713-735.

SIDER, David. 2009. The special case of Herculaneum, in: Roger S. BAGNALL (ed.), *The Oxford Handbook of Papyrology*, Oxford, Oxford University Press, p. 303-319.

SKUT, Wojciech, Brigitte KRENN, Thorsten BRANTS & Hans USZKOREIT. 1997. An annotation scheme for free word order languages, in: *Fifth Conference on Applied Natural Language Processing*, Washington, DC, USA, March–April 1997, p. 88-95.

WILSON, Mark. 1997. Ancient Egyptian language discussion list, http://www.rostau.org.uk/AEgyptian-L/, 1997. Accessed 2011-10-14.

WITTENBURG, Peter, Hennie BRUGMAN, Albert RUSSEL, Alex KLASSMANN & Han SLOETJES. 2006. ELAN: A professional framework for multimodality research, in: *LREC 2006: Fifth International Conference on Language Resources and Evaluation, Proceedings*, Genoa, p. 1556-1559.

ŽÁBA, Zbynek. 1956. *Les Maximes de Ptahhotep*, Prague, Éditions de l'Académie Tchécoslovaque des Sciences.

Hieroglyphic Text Processors, Manuel de Codage, Unicode, and Lexicography*

Roberto B. GOZZOLI

Mahidol University

1. INTRODUCTION

This paper represents a sort of junction between different ideals. It addresses the general insularity into which Egyptology has fallen, and more precisely the niche Egyptologists have excavated for themselves within the small group of scholars specializing in the ancient world.

Ancient Egypt is certainly one of the most fascinating cultures of the world. But the specialists of its study have not contributed to the creation of a climate of collegiality in their discipline. Each scholar secludes himself within his own specialization, defeating the spirit of collaboration quite typical of the early twentieth century. Here I do not have to look further than the *Wörterbuch*, which saw scholars from different countries contribute to the different entries.[1] Collegiality such as this is needed now more than ever in Egyptology, as new challenges arise and new directions are forged, particularly with the rise of computer-aided research.

2. HIEROGLYPHIC TEXT PROCESSORS

In the beginning, there was Glyph, a program created by Jan Buurman that is generally regarded as the beginning of computer-aided hieroglyphic typesetting. It was not the only product available and others were already in business by the end of 1960s and beginning of 1970s.[2] Glyph development at the end of 1980s was connected with the growing importance of computers in Egyptology, demonstrated by the publication of the Manuel de Codage in 1988.[3] It was a sort of transition from the typewriter to the word processor and, for the hieroglyphs, from manual to digital typesetting. Not long before this, most hieroglyphs had to be handwritten once the final copy was ready to print, so the possibility of having them computer generated was certainly a big step forward.

[*] I am grateful to Stéphane Polis for the help in delivering the original poster to the audience, due to my inability to be present at the workshop. In order to avoid lengthy footnotes with website addresses, I cite them in the bibliography. I must also apologize for my paper in many respects, perhaps the greatest of which is that, as an historian giving his impressions of computers and Egyptology, I have a limited perspective: how technology can or should aid my research. Thus, I am presenting a vision that may be not highly technical.

[1] See Erman & Grapow 1926-1963. Apart from the Berlin Academy, Alan Gardiner and James Breasted contributed to the project as Adolf Erman's pupils.

[2] See the timeline of computer development within Egyptology as presented in the Rosette webpage (referred in the bibliography) as well as in the current volume.

[3] Buurman *et al.* 1988.

In 1986, Dirk van der Plas was one of the earliest Egyptologists using Glyph for his publication of the Hymn to the Nile.[4] At that time, Glyph was not alone: Peter der Manuelian used a specialised font for his book on Amenhotep II.[5]

While Glyph (still on the DOS platform) continued to prosper until the beginning of the 1990s, other systems and methods were available: at the Metropolitan Museum of Art in New York, James Allen created his own font to be manipulated using CorelDraw.[6] In 1991 versions of hieroglyphic typesetting on Atari were also available and used in Bernadette Menu's Grammar and Dictionary.[7]

The Apple Macintosh world was not left behind, as they had their own version of Glyph, named MacScribe. While I cannot talk specifically about its early developments as I had no Apple computer at that time, the coding for the Mac version was the same as the Windows version. Both shared the Manuel de Codage, implementing in full the conventions. Indeed, some would say that the Manuel de Codage was built on top of Glyph functionality.[8] The graphic capabilities of the operating system made possible some drag and drop operations.

These hieroglyphic text processors were quite divergent, as there were many operating systems at the time, and Windows was not yet fully developed.

The explosion of Windows, in particular 3.1, and then Windows 95 and 98 saw the transition of Glyph to Winglyph (fig. 1). Version 1.2 came out in 1996, and the final release is dated 2001. These versions, as well as the Mac counterpart, were extended by *Hieroglyphica*, a library of Late Period and Ptolemaic signs, mostly developed by Jochen Hallof, with version 1 released contemporary with WinGlyph in 1996 and version 2 coming out in 2001. In fact, the 2001 version was nothing more than an updated *Hieroglyphica*, and there was no change for WinGlyph itself.

There were (and are) some odd things to be noted about WinGlyph. The line markers (a vertical line with the line number at the top) could be printed, but whenever you tried to export the text to Word for instance, they disappeared from the pasted version. It has never been possible to rectify this irregularity. The explanation in the manual that WinGlyph was designed to produce texts and then print directly has never been satisfactory in any case, as many Egyptologists used it to produce short texts or words to be inserted into and mixed with non-hieroglyphic text, for instance in a textual commentary. It is also odd that a sign editor has never been released. This deficit had some disastrous consequences on some of my work as an undergraduate student, as some hieroglyphs were not present in the first version of Winglyph. At the time I was conducting research on Psammetichus II, and several signs used in some of the relevant stelae were not part of the hieroglyphic set.[9] The absence of a hieroglyphic editor was undoubtedly quite substantial.

Whatever the case, Winglyph and MacScribe were the tools for any book with hieroglyphs printed in the period between 1988 and 2005. You can recognize that the font is based on the style of Gardiner's *Egyptian Grammar*.[10]

The spread of WinGlyph during the 1990s was encouraged by a series of initiatives headed by the CCER, including the Multilingual Egyptological Thesaurus,[11] and since then Glyph has essentially

[4] Cf. van der Plas 1986.

[5] Der Manuelian 1987 for the book on Amenhotep II.

[6] Kind information by James Allen (2 October 2011; email).

[7] Menu 1989; 1990.

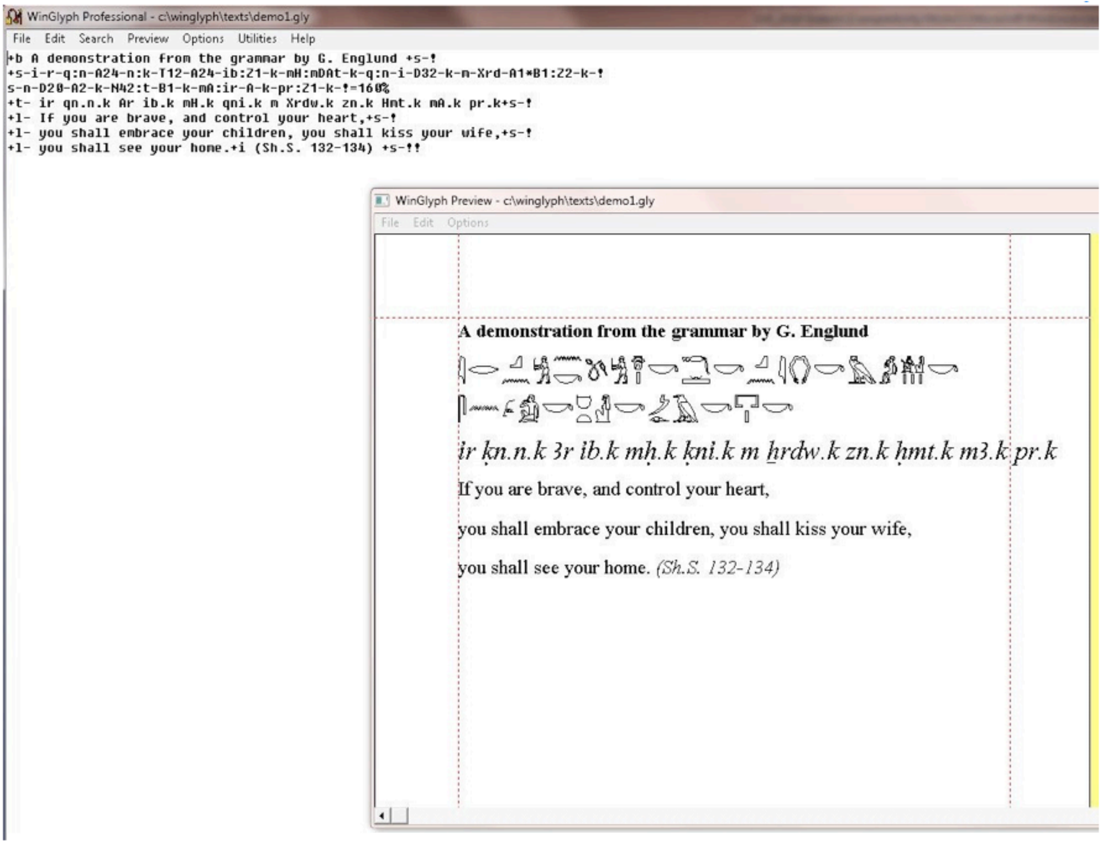
[8] I refer to Nederhof 2002.

[9] This work resulted in a thesis, which is now being updated and revised for publication.

[10] The basic signs in WinGlyph from Gardiner's library were more simplified than those in the MacScribe version, for reasons that perhaps should be investigated.

[11] For the Thesaurus, see van der Plas (1996) and www.ccer.nl/apps/thesaurus/index.html.

been the base of any hieroglyphic text processor on PCs, as MacScribe has been for Mac computers. But Winglyph now belongs to the past, as any further development is ceased and the CCER shut down.[12]


WinGlyph Professional - c:\winglyph\texts\demo1.gly
File
Edit
Search
Preview
Options
Utilities
Help
WinGlyph Preview - c:\winglyph\texts\demo1.gly
File
Edit
Options

```
+b A demonstration from the grammar by G. Englund +s-!
+s-i-r-q:n-A24-n:k-T12-A24-ib:Z1-k-mH:mDAt-k-q:n-i-D32-k-m-Xrd-A1*B1:Z2-k-!
s-n-D20-A2-k-N42:t-B1-k-mA:ir-A-k-pr:Z1-k-!=160%
+t- ir qn.n.k Ar ib.k mH.k qni.k m Xrdw.k zn.k Hmt.k mA.k pr.k+s-!
+l- If you are brave, and control your heart,+s-!
+l- you shall embrace your children, you shall kiss your wife,+s-!
+l- you shall see your home.+i (Sh.S. 132-134) +s-!!
```

A demonstration from the grammar by G. Englund

ir ḳn.n.k ꜣr ib.k mḥ.k ḳni.k m ẖrdw.k zn.k ḥmt.k mꜣ.k pr.k

If you are brave, and control your heart,

you shall embrace your children, you shall kiss your wife,

you shall see your home. *(Sh.S. 132-134)*

Figure 1. Winglyph 2.0

As I look at the past, especially almost two decades ago, computers and the Humanities were slow to move on early developments. There was lot of enthusiasm, surely a lot of inexperience, but at the same time it was a period where a focused computer centre in Egyptology could have pushed in new directions.

It was a pioneer era, when young or not-so-young Egyptologists sat down and tried to make use of the computer for their researches. The multiplicity of hieroglyphic text processors of that time, sometimes used only by one researcher, is proof of this. And the fact that that for six years there has been a section in the *Bulletin de la Société d'Égyptologie de Genève* dedicated to computer projects in Egyptology demonstrates such interests in it.[13] The establishment of CCER should also be seen from this point of view: an enthusiastic approach to computers, which was essentially a university project and that ran well as long as funding was provided.[14]

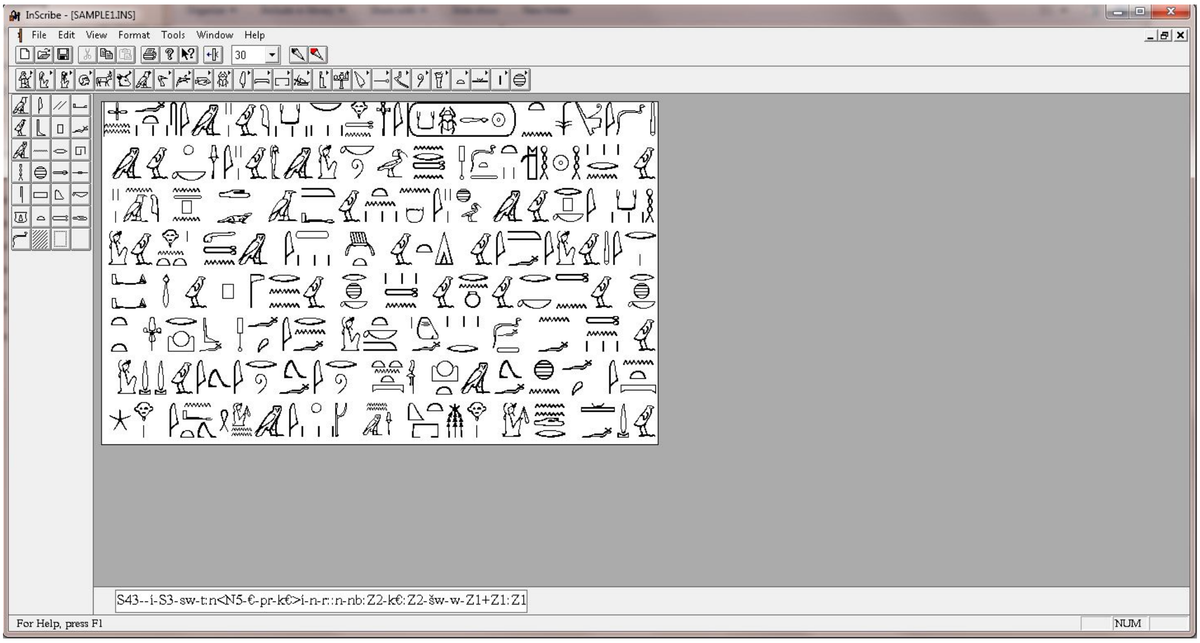
As it is now time to say goodbye to Glyph and its incarnations, what can be said of the experience itself is that Glyph was the first large scale implementation of a hieroglyphic text processor.

[12] At the end of September 2010, a letter by Dirk van der Plas appeared on the various mailing lists (EEF, Agade) as well as on the CCER website, saying that the CCER shop was going to be decommissioned by the end of the year. The website itself would be maintained by Hans van den Bergh, to be used as part of the history of Computers and Egyptology.

[13] Published between 1995 and 2001, and now online (see Websites list).

[14] While I will return to the CCER as an institution later on, the absence of any development on WinGlyph from 1996 has been a serious mistake by the CCER itself, with the benefit of hindsight. There was feedback from users and the limitations of the project were acknowledged but never rectified. I feel this to be a sort of contradiction to what a scientific institution should be.

Of the group of hieroglyphic text processors available during the 1990s, and with the obvious exclusion of MacScribe, only one is still in production, Inscribe. First released by Bob Richmond in 1994 (fig. 2), Inscribe was more click-based, as the sign palette was more immediate than in WinGlyph. The basic coding of the signs was the same as Winglyph, but how the signs were grouped or superimposed was not each time the same, so there would be compatibility problems (fig. 3). I believe that the deviation from the Manuel de Codage has been corrected in the later versions.

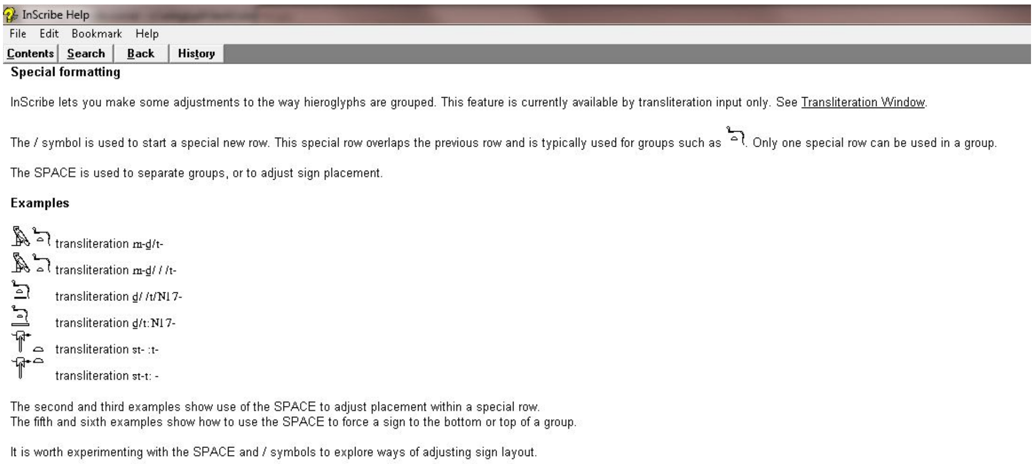


S43--i-S3-sw-t:n<N5-€-pr-k€>i-n-r::n-nb:Z2-k€:Z2-šw-w-Z1+Z1:Z1

For Help, press F1

Figure 2. Inscribe 1.0

Version 1 of Inscribe was in widespread use until 2004, when a new version came out. While I cannot talk about those later developments, which seem to have stalled, at least for the last release, some of the initial implementations presented a better integration with Microsoft Word. OLE (Object Linking and Embedding) integration was supplied, so you could click in the Hieroglyphic text in Word, thus opening Inscribe and modifying the hieroglyphs. For WinGlyph, everything was (and still is) cut-and-paste. As far as I know, the major Egyptological institution using Inscribe is the Griffith Institute in Oxford. Bob Richmond and Mike Everson however have been amongst the main supporters of the implementation of Egyptian hieroglyphs in Unicode, someone would say for commercial reasons.



Special formatting

InScribe lets you make some adjustments to the way hieroglyphs are grouped. This feature is currently available by transliteration input only. See Transliteration Window.

The / symbol is used to start a special new row. This special row overlaps the previous row and is typically used for groups such as . Only one special row can be used in a group.

The SPACE is used to separate groups, or to adjust sign placement.

Examples

transliteration m-ḏ/t-
transliteration m-ḏ/ / /t-
transliteration ḏ/ /t/N17-
transliteration ḏ/t:N17-
transliteration st- :t-
transliteration st-t: -

The second and third examples show use of the SPACE to adjust placement within a special row.
The fifth and sixth examples show how to use the SPACE to force a sign to the bottom or top of a group.

It is worth experimenting with the SPACE and / symbols to explore ways of adjusting sign layout.

Figure 3. Inscribe 1

The promises of Inscribe 3 seem to get closer to a hieroglyphic Word processor (capital letter intentional), supplying many of the same characteristics you can find in Microsoft Word 2007 (justification, spacing and so forth).

While all those improvements are praiseworthy, some more recent products give a better implementation of "real" hieroglyphs. The first product of the new millennium I should mention is VisualGlyph (fig. 4), developed by Gunther Lapp. Lapp's approach still follows some of the sign coding of *Hieroglyphica*, which is not followed by others. Moreover, the coding of the composition of groups of signs does not match any other hieroglyphic software. Instead of the colon (:), the superimposition of two signs is achieved by the forward slash (/). But the software has some very interesting features: first of all, the hieroglyphs are True Type fonts, so any missing hieroglyphs can be added with font maker software. The other major advantage is certainly the ability to put the hieroglyphs in different positions, thus really matching the original disposition of the text on a papyrus or a vase for example. Version 1 came out in 2003, while version 2 is dated 2004. In an email (3 June 2010), I have been informed that any further development will be a DotNet version.[15]

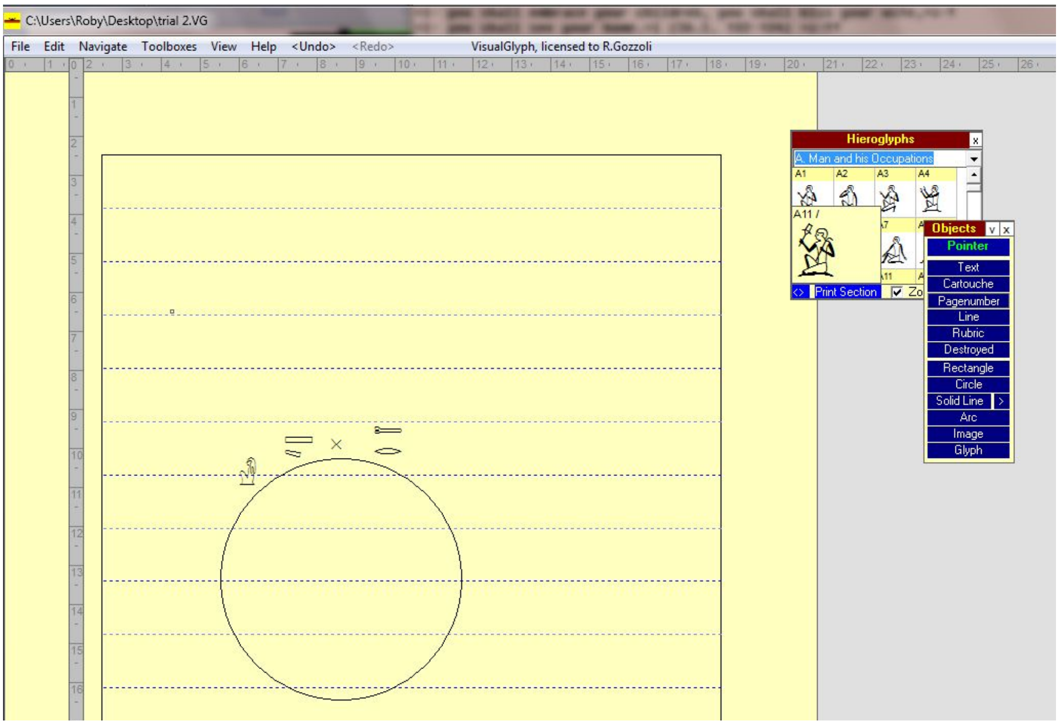


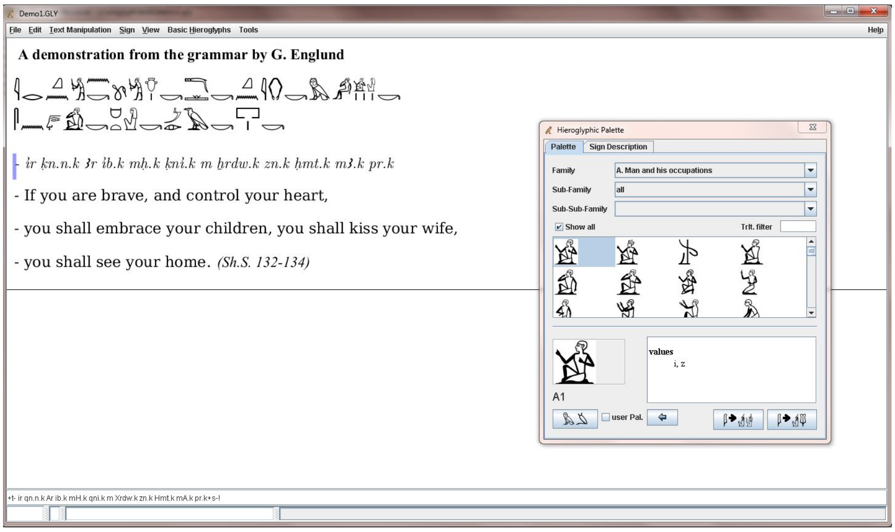
Figure 4: VisualGlyph 2.0

Another major product and in my opinion the successor to WinGlyph is JSesh (fig. 5), developed by Serge Rosmorduc, who created other hieroglyphic product such as TkSesh and HieroTeX.[16] Version 2.0 alpha was released in 2006, according to the JSesh website.

The most interesting aspect is that JSesh is able to read the Manuel de Codage fully, thus permitting the translation of old WinGlyph files. The opposite is not allowed, as JSesh contain extensions to the Manuel.

[15] DotNet (.Net) framework is a software framework used to maintain interoperability between old and new software application, independently from the hardware environment.

[16] TkSesh is a hieroglyphic database system, developed during the 1990s (webperso.iut.univ-paris8.fr/~TKSESH/IEXII/tksesh.html). HieroTeX is a computer package through which hieroglyphs can be inserted into a LaTeX scripting environment (http://webperso.iut.univ-paris8.fr/~rosmord/archives/).



A demonstration from the grammar by G. Englund

- ỉr ḳn.n.k ȝr ỉb.k mḥ.k ḳnỉ.k m ẖrdw.k zn.k ḥmt.k mȝ.k pr.k

- If you are brave, and control your heart,

- you shall embrace your children, you shall kiss your wife,

- you shall see your home. *(Sh.S. 132-134)*

Family: A. Man and his occupations
Sub-Family: all
Sub-Sub-Family:
☑ Show all
Trlt. filter:

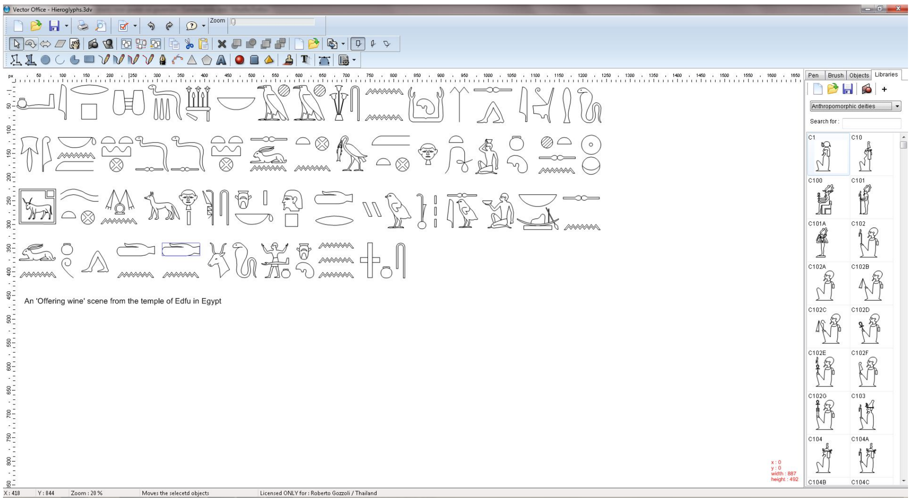
values
i, z

+t- ir qn.n.k Ar ib.k mH.k qni.k m Xrdw.k zn.k Hmt.k mA.k pr.k+s-!

Figure 5. JSesh 4.3

The product can only write linear text, and so looks quite similar to Winglyph or Inscribe. It contains a sign editor and it is the first hieroglyphic text processor to be offered completely free. As it is written in Java, it runs on different operating systems without problems. Text can be entered via the sign palette or through the normal code typing.[17] The great value of JSesh is that it is still in full development and new releases are quite frequent.[18] The fact that the software is developed in cooperation with IFAO is also a guarantee that Egyptologists will be consulted in the design process. Moreover, Serge Rosmorduc is both a software developer and a lecturer in Egyptology at the École Pratique des Hautes Études, and this ensures a scholarly understanding of the relevant issues.

Another hieroglyphic text processor recently developed is VectorOffice[19] (fig. 6), produced by Jan-Peter Graeff. The product itself is very interesting as it joins together a hieroglyphic editor within an interface that allows for graphical manipulation. A 3D object can be created inside VectorOffice itself, avoiding its export to different graphic software. The format for export is Windows Metafile, which means the export is still limited to the figure as block, different to JSesh in this respect, though it allows the manipulation of the position of the hieroglyphs within Microsoft Word for instance.



Anthropomorphic deities

An 'Offering wine' scene from the temple of Edfu in Egypt

X : 418 | Y : 844 | Zoom : 20 % | Moves the selecetd objects | Licensed ONLY for : Roberto Gozzoli / Thailand

Figure 6. VectorOffice 2011

[17] Some purists do not like the fact that the signs are not as similar to those in Gardiner's *Egyptian Grammar* as all the other products. While the objection is essentially valid, in time such 'nuisance' will be completely disregarded.

[18] The latest release to date (14 September 2012) is version 6.0.0 (see jsesh.qenherkhopeshef.org).

[19] With the support of the Edfu Project team (Dieter Kurth).

The last program to be discussed is Hieroglyphica (fig. 7). Developed by Maxim Panov (Novosibirsk, Russia), it is a hieroglyphic word processor similar to Inscribe in some respects, but completely independent from it.

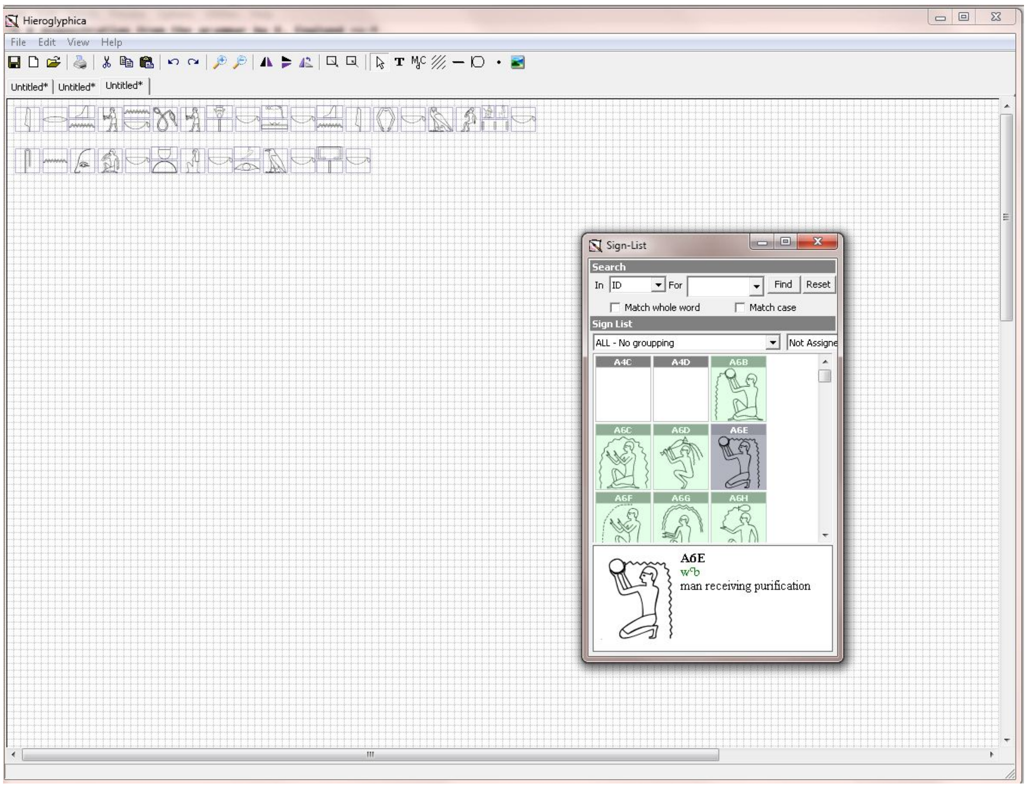


Figure 7. Hieroglyphica

Having listed all those products, it has to be noted that only JSesh and Inscribe 2004 are able to read the files as `.gly` format, the WinGlyph format. All the others save the documents in their own specific format. Therefore, no file exchange is possible between the different products. If I send my VisualGlyph file to a colleague in JSesh, I will be obliged to send it as a PDF or attached to a Word file, but another program will be unable to make a correction to the file itself. Of course, using Adobe Photoshop or any graphical software will always be able to manipulate the image itself, but I still have to modify the original file anyway. Moreover, if I change software product in the middle of a project for any reason, the incompatibility of WinGlyph with VisualGlyph will mean that I have to retype small or large portions of hieroglyphic text. Certainly this is not the most attractive option for pressing deadlines or very busy scholars.

I will close this historical overview of the hieroglyphic texts editors with three suggestions:

- It would be useful to develop some kind of translator between the various formats, i.e. a self-standing format that I can open in any tool I use for typing my hieroglyphs.
- As I have discussed above, all the actual software products have their pros and cons, but I would argue in favour of a cross platform product that can be used independent of any current operating system (given the rate of change of such things).
- As computer software in Egyptology is actually independent from major software companies, I believe that finding a common format able to read the existing ones and produce a generic output would only require some goodwill by anyone involved in those hieroglyphic typewriting software projects. As scholars and departments are getting short-funded, this

venture will ensure that anything done now in 2011 will be still a valid standard for 2020 for instance — I hope for longer than that — in spite of the inevitability of great technological change.

3. Manuel de Codage

The Manuel de Codage certainly has a long history behind it. Without belabouring this, however, it can be said that the Manuel has been the standard since 1988.[20] While in Egyptological terms this certainly cannot be considered as a past so distant, in computer terms it is like being in a different era.

From a computer point of view, the limitations of the Codage are evident: the grouping of the signs by (*), (:) and (,) has been useful for printing results, but certainly meaningless in computer terms. As remarked by Mark Nederhof,[21] the positioning of signs is not considered by the Manuel, and even the scaling of the hieroglyphs does not make any sense for any hieroglyphic software, with the obvious exclusion of WinGlyph. In addition, the problem of sign positioning outside the square unit in which the signs should be written could never be resolved. Hence, if I encode hieroglyphs in a circle, like those on a bowl, the positioning cannot be described, as there is no general implementation. I can do it, but I cannot replicate it outside that specific software.

Egyptologists do not seem to have cared about it, but their reasons are quite understandable. Up until now, hieroglyphic typing has only been seen as a tool to put Egyptian hieroglyphs into print publications. The print version was the final product. The idea of a common repository of texts has not yet filtered through. If I ask for some text from a colleague for a common project, at the present times he needs to send me a PDF file of the hieroglyphs, if not an ink drawing of it, and this is all I can get. If we share the same hieroglyphic word processor, of course, a file can also be sent with the encoded text.

With a new implementation of the Manuel following some specific directives, however, it would be enough to send me a codified version, and translate it with specific software. Thus I can have a computer generated hieroglyphic text, including grouping and positioning. And this will be completely independent from the actual software used, as long as the program I am using is able to read and understand the underlying code.

There are many limits to the Manuel de Codage that the actual software goes much beyond. And yet, a new standard is still far from being defined. Nederhof's proposal in 2002, and Rosmorduc's reworking have yet to find widespread usage for it: some papers in this volume and elsewhere have focused on XML, so there is no need to go on any further about it.[22] My claim at this point is that a new committee should be set up for a new version of the Manuel, to try to create a new set of rules applicable to the software we have now, to forecast any possible developments, and to take into consideration sign positioning within the myriad of texts available and known until now.

4. Unicode fonts

Recent implementations of Unicode standards have demonstrated some integration of Egyptology within the more general linguistic and academic community. For the transliteration, we are quite satisfied with the signs we have — of course with the exception of *yod* — but future software development will improve the situation.[23]

[20] Buurman *et al.* 1988.

[21] Nederhof 2002.

[22] See Nederhof in this volume.

[23] See the discussion by Everson & Richmond 2008. Egyptological alef and ayin have been integrated into Unicode 5.1 and the hieroglyphic signs with Unicode 5.2. Otherwise, another option is to change the general habit in favour of using *j* instead.

As now more than a thousand Egyptian hieroglyphs are actually part of Unicode, it is remains a fact that, at the moment, only two fonts are available: Mark Nederhof's Gardiner Font and George Douros' font.[24]

For the hieroglyphic signs, we have now more than what Alan Gardiner proposed in his Grammar more than half a century ago.[25] The Unicode list is quite comprehensive. Obviously it will never be a replacement of WinGlyph's *Hieroglyphica* for instance, but it should be more than acceptable for any Middle Egyptian texts.

While classical languages such as Latin and Ancient Greek may have the luxury of variants for some signs, as their alphabet is numerically small, the vast number of hieroglyphs stretching over various periods prohibits such an enterprise — if there should ever be a reason for it.

The general Egyptological community has been quite slow to adopt the new implementations. In fact, the adoption of Egyptian hieroglyphs and Unicode was a sort of commercial effort by Michael Everson in cooperation with Bob Richmond of Saqqara software.[26] It is disappointing that two years after Unicode 5.2 we are still waiting for Inscribe 3 as the software *par excellence* that should be able to use those fonts. I understand that the Egyptological community was quite sceptical about the results and in disagreement with the commercialization as given in the Saqqara proposal. As for its actual usage now in 2011, the hieroglyphs in Unicode can be used as a sort of nice decoration within a normal text, but it is not yet possible to render a real hieroglyphic text as written on ancient Egyptian monuments.

Whatever the case, the technology is available and Egyptologists should try to make a good use of it. While it is understandable that it has almost no practical function at the moment, a special committee should be ready to recognize future developments and implement applications as soon as they appear.

5. Hieroglyphic databases

One of the foci of this paper concerns lexicography. Textual databases have been in use by various Egyptologists in their researches and textual analysis in search for parallels has been one of the major relevant topics.

At the moment, three public outputs are known to me.[27] One is *Ramses*, from the University of Liège, for the study of Late Egyptian.[28] The other one is SESCH by Eberhard Holzhäuer, which is supported by the University of Marburg.

There is also a private product, the program *Corpus*, developed by Maxim Panov, the only program I have effectively tested. It is quite accurate, and it does what it promises (though I am not acquainted well enough with it to appraise its strengths and weaknesses).

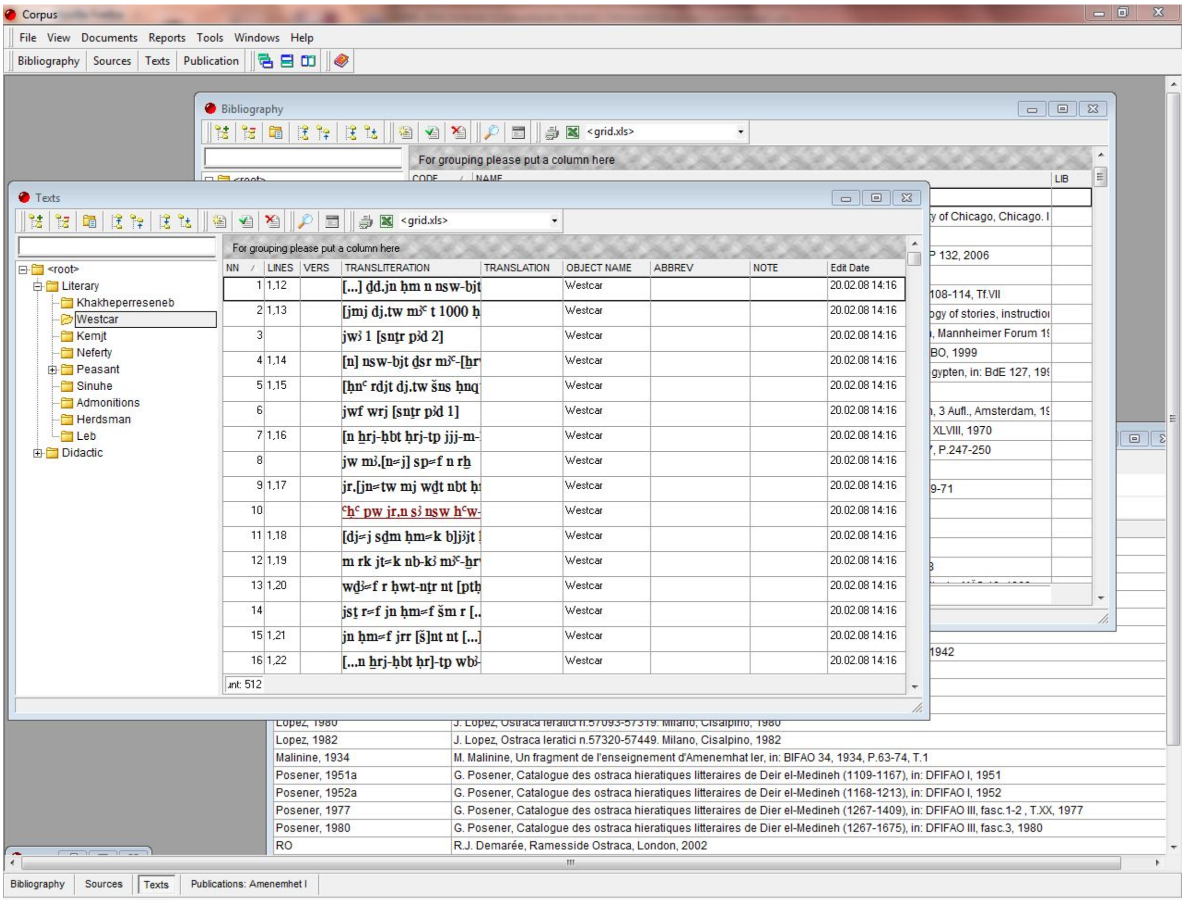
[24] New Gardiner font: www.cs-st-andrews.ac.uk/~mjn/egyptian/fonts/newgardiner.html; Aegyptus font: greekfonts.teilar.gr/~g1951d.

[25] Gardiner ^{3}1957.

[26] See Everson 2006.

[27] As I can testify their existence, but not their functions, the discussion here is admittedly quite limited.

[28] See the database resource list (bibliography) and the relevant papers in the current volume.


Corpus
File View Documents Reports Tools Windows Help
Bibliography Sources Texts Publication
Bibliography
<grid.xls>
For grouping please put a column here
CODE NAME LIB
Texts
<root>
Literary
Khakheperreseneb
Westcar
Kemjt
Neferty
Peasant
Sinuhe
Admonitions
Herdsman
Leb
Didactic
Bibliography Sources Texts Publications: Amenemhet I

For grouping please put a column here

NN	LINES	VERS	TRANSLITERATION	TRANSLATION	OBJECT NAME	ABBREV	NOTE	Edit Date
1	1,12		[...] ḏd.jn ḥm n nsw-bjt		Westcar			20.02.08 14:16
2	1,13		[jmj dj.tw mꜣꜥ t 1000 ḥ		Westcar			20.02.08 14:16
3			jwꜣ 1 [snṯr pꜣd 2]		Westcar			20.02.08 14:16
4	1,14		[n] nsw-bjt ḏsr mꜣꜥ-[ḫr		Westcar			20.02.08 14:16
5	1,15		[ḥnꜥ rdjt dj.tw šns ḥnq		Westcar			20.02.08 14:16
6			jwf wrj [snṯr pꜣd 1]		Westcar			20.02.08 14:16
7	1,16		[n ḫrj-ḥbt ḥrj-tp jjj-m-		Westcar			20.02.08 14:16
8			jw mꜣ.[n=j] sp=f n rḫ		Westcar			20.02.08 14:16
9	1,17		jr.[jn=tw mj wḏt nbt ḥ		Westcar			20.02.08 14:16
10			ꜥḥꜥ pw jr.n sꜣ nsw ḫꜥw-		Westcar			20.02.08 14:16
11	1,18		[dj=j sḏm ḥm=k b]jꜣjt		Westcar			20.02.08 14:16
12	1,19		m rk jt=k nb-kꜣ mꜣꜥ-ḫr		Westcar			20.02.08 14:16
13	1,20		wḏꜣ=f r ḥwt-nṯr nt [ptḥ		Westcar			20.02.08 14:16
14			jsṯ r=f jn ḥm=f šm r [..		Westcar			20.02.08 14:16
15	1,21		jn ḥm=f jrr [š]nt nt [...]		Westcar			20.02.08 14:16
16	1,22		[...n ḫrj-ḥbt ḥr]-tp wbꜣ-		Westcar			20.02.08 14:16

ınt: 512

Lopez, 1980	J. Lopez, Ostraca Ieratici n.57093-57319. Milano, Cisalpino, 1980
Lopez, 1982	J. Lopez, Ostraca Ieratici n.57320-57449. Milano, Cisalpino, 1982
Malinine, 1934	M. Malinine, Un fragment de l'enseignement d'Amenemhat Ier, in: BIFAO 34, 1934, P.63-74, T.1
Posener, 1951a	G. Posener, Catalogue des ostraca hieratiques litteraires de Deir el-Medineh (1109-1167), in: DFIFAO I, 1951
Posener, 1952a	G. Posener, Catalogue des ostraca hieratiques litteraires de Deir el-Medineh (1168-1213), in: DFIFAO I, 1952
Posener, 1977	G. Posener, Catalogue des ostraca hieratiques litteraires de Dier el-Medineh (1267-1409), in: DFIFAO III, fasc.1-2 , T.XX, 1977
Posener, 1980	G. Posener, Catalogue des ostraca hieratiques litteraires de Dier el-Medineh (1267-1675), in: DFIFAO III, fasc.3, 1980
RO	R.J. Demarée, Ramesside Ostraca, London, 2002

Figure 8. Corpus 2.0

Many other database tools are certainly available for various projects; it is enough to check the Egyptology pages kept by Nigel Strudwick.[29] Yet again there is the problem of the software architecture used and the ability to share data. How can data exchange be possible through different formats and platforms?

Moreover, the disappearance of Peter Jurgens' Access database of Coffin Texts is symptomatic of two other issues:[30]

– If the work is still valid, as the fact that someone is still searching for it seems to imply (see previous note), why should the Egyptologists' community be impoverished by its disappearance? Perhaps researchers at Göttingen have moved away from an interest in the Coffin Texts, but a nice piece of work has nonetheless disappeared. Considering that the archaeological departments are still full of students, this work is still valuable, even if never actually published. Years of someone's life and effort should not be thrown away or relegated to dusty bookshelves before being packed away into obscurity. Pursuing this topic further would lead to the larger issue of the dissemination of research in general, but I will keep the discussion to computers and Egyptology. Why should useful computer projects (database, corpus, whatever) be fated to oblivion?

– The second aspect is obviously software obsolescence. The fact that old Access files are not readable by newer Access versions demonstrates the myopic vision of Microsoft, only centred toward selling their products, without thinking about long term customer satisfaction. Yet,

[29] See www.fitzmuseum.cam.ac.uk/er.

[30] This database was written in Microsoft Access 97, which no longer functioned in Access 2007. The original database is no longer on the Göttingen University, Egyptological Department webpage (cf. email from Peter Robinson in the EEF List, 11 September 2010). Webpage changes have most probably prompted the cleaning of old material.

> this is again a problem for Egyptologists, as this means that even for our daily activities of research, we may find that some old research done 15 years ago and still kept in our computers is really worth nothing, in spite of all the money spent for last generation computers and software.

We are studying a dead civilization, which was able to maintain itself for 3,000 years and yet many of the tools we use to understand its language and civilization are doomed to be obsolete 20 years after their release or production. Therefore, there is the need of a working group dealing this problem as well as a computer archive to keep an updated electronic archive of any application within Egyptology. This central archive would be able to keep an electronic copy of the software material produced by a scholar during his/her research and keep it updated so as to make it accessible long after the research project is finished.

There are certainly issues with this idea: legal (software and intellectual copyright), financial (the software will be needed to keep it updated, so it will cost) and personal (scholars sometimes prefer not to make their work freely accessible, possibly fearing for plagiarism). But I do not see all these points as impossible to overcome: economy should be the magic word. Why should I waste energy reduplicating someone else's work? Research should be based on how I can use my predecessors' fruitful work to make a further step forward.

Finally, there is the issue of recognition. If for instance I write about the historical texts of the Middle Kingdom, I would be grateful to use some previous research or computer databases produced by my colleagues, and I will mention such contributions in the same way I would quote a written publication.

6. Conclusions

This paper addresses some current issues and appears nostalgic in tone. But I would conclude my paper suggesting an idea for the future: the establishment of a new CCER. I understand that this is certainly difficult due to the financial issues of research institutes. Yet the financial problems should encourage Egyptologists to abandon isolationism and competitive feelings and really fight for common goals. All we need is an established working group for computers and Egyptology, a mix of Egyptologists and computer specialists who are able to examine the actual state of the art in computer hardware and software, and realize appropriate products.

A communal project I also feel is necessary is an Egyptological software system, where different modules (lexicographical database, bibliographic database, hieroglyphic text processor, word processor) can be developed, all of them focused within the specific requirements of Egyptological community, as well as anyone working with ancient or modern texts. Leaving aside the hieroglyphic processing, already discussed above, at the moment we are at the mercy of Microsoft Word, Endnotes and Access (or FileMakerPro) for our needs, with exceptions for some specific applications. All the software mentioned has problems in rendering hieroglyphs and/or indexing ancient Egyptian words, due to the particularities of sign ordering, and this provides another motivation for an Egyptology-specific text processor. As the workstation could be modular, it should be quite easy to complete each part, without having a full product immediately.

As remarked many times it depends on only one requirement: sharing. If the idea of collegiality prevails, then in due time Egyptologists will have their own scientific tools as well as a common platform — and common data — within our community. If not, everything will continue as it is now: independent projects + independent ideas = duplication of material → waste of time. I let the reader choose the way.

BIBLIOGRAPHY

BUURMAN, Jan, Nicolas GRIMAL, Michael HAINSWORTH, Jochem HALLOF & Dirk VAN DER PLAS. 1988. *Manuel de codage des textes hiéroglyphiques en vue de leur saisie sur ordinateur*, Paris, Institut de France.

DER MANUELIAN, Peter. 1987. *Studies in the Reign of Amenophis II*, Hildesheim (= *HÄB* 26).

ERMAN, Adolf & Hermann GRAPOW (eds.). 1926-1963. *Wörterbuch der ägyptischen Sprache*, 7 vol., Leipzig-Berlin.

EVERSON, Michael. 2006. Sources for the encoding of Egyptian Hieroglyphs, in: http://std.dkuug.dk/jtc1 /sc2/wg2/docs/n3182.pdf.

EVERSON, Michael & Bob RICHMOND. 2008. EGYPTOLOGICAL YOD and Cyrillic breathings, in: http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3382.pdf.

GARDINER, Alan H. ^{3}1957. *Egyptian Grammar: Being an Introduction to the Study of Hieroglyphs*, Oxford, Griffith Institute.

MENU, Bernadette. 1989. *Petite grammaire de l'égyptien hiéroglyphique à l'usage des débutants*, Paris.

—. 1990. *Petit lexique de l'égyptien hiéroglyphique à l'usage des débutants*, Paris.

NEDERHOF, Mark-Jan. 2002. A Revised Encoding Scheme for Hieroglyphic, in: http://www.cs.st-andrews.ac.uk/ ~mjn/egyptian/res/ie2002.pdf

VAN DER PLAS, Dirk. 1986. *L'hymne à la crue du Nil*, Leiden, Nederlands Instituut voor het Nabije Oosten (= *Egyptologische Uitgaven* 4/1-2).

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

WEBSITES

History of Computers and Egyptology

Computer and Egyptology history: http://projetrosette.info/assets/echange/Historique_InE.xls
Roger Monfort's history: http://raymond.monfort.free.fr/1.html
Ressources Égyptologiques Informatisées: http://www.segweb.ch/resso.htm

Hieroglyphic processors

Hieroglyphica: http://www.hieroglyphica.com/hieroglyphica.php
JSesh: http://www.jsesh.qenherkhopeshef.org
MacScribe: http://www.macscribe.com
VectorOffice: http://www.hornet-sys.com/VectorOffice.html
VisualGlyph: http://aegyptologie.unibas.ch/online-tools/visualglyph-for-pc/
Winglyph and Hieroglyphica: www.ccer.nl

Manuel de Codage

Mark-Jan Nederhof: http://www.cs.st-andrews.ac.uk/~mjn/egyptian/
Serge Rosmorduc: http://www.iut.univ-paris8.fr/~rosmord/HieroEncoding/DTD/
Summary: http://www.catchpenny.org/codage/

Fonts

New Gardiner: http://www.cs.st-andrews.ac.uk/~mjn/egyptian/fonts/newgardiner.html
Aegyptus: http://users.teilar.gr/~g1951d/
HieroTeX: http://webperso.iut.univ-paris8.fr/~rosmord/archives/

Databases

Ramses Project: http://www.egypto.ulg.ac.be/docs/Ramses_NewResearchTool.pdf;
http://www.egypto.ulg.ac.be/Manuel_Ramses.pdf (for the references)

SESCH: http://www.sesch-projekt.de

TkSesh: http://webperso.iut.univ-paris8.fr/~TKSESH/IEXII/tksesh.html

TUSTEP: http://www.tustep.uni-tuebingen.de/tustep.html

Corpus: http://www.hieroglyphica.com/download.php

The Manuel de Codage Encoding of Hieroglyphs Impedes Development of Corpora[*]

Mark-Jan NEDERHOF

University of St Andrews

1. INTRODUCTION

The Ancient Egyptian hieroglyphic writing system has a number of properties that set it apart from most other modern and ancient writing systems (Daniels & Bright 1996). One is that the pictographic aspect was maintained throughout its history. Stylisation and abbreviation of signs have played a much smaller role than in the cases of for example Akkadian cuneiform or Chinese. Whereas hieratic can be seen as a cursive form of hieroglyphic, the latter was never replaced by the former, and they influenced one another throughout history. Despite this moderate degree of character stylisation, there was no limit on the number of signs that could be used, and large variation can be observed in their exact appearance.

A second aspect of hieroglyphic writing that sets it apart is a particular form of aesthetics, including a desire to divide the available surface in a way pleasing to the eye, avoiding large empty spaces. Thus, two signs with large height and small width could be placed one next to the other, and two signs with small height and large width could be placed one below the other. This is however by no means the only way of placing signs relative to one another. Frequently, the empty space in the corner of one sign is used to harbour a second sign of small size. One sign can also be placed inside another or two signs can be positioned one overlapping the other, especially if there is a linguistic connection (e.g. collocation) between the words the two signs represent.

Because Ancient Egyptian hieroglyphs seem to form an exceptional case within the wide range of the world's writing systems, it is not *a priori* clear that common technical solutions that have been devised for processing other writing systems are also suitable to hieroglyphs. A good illustration is perhaps the arduous process that has led to the inclusion of hieroglyphs in Unicode.

The first proposal to include Ancient Egyptian was proposal N1637, undertaken by Michael Everson (1997). This was based on the sign list from Gardiner (^{3}1957) and comprised 761 signs, together with operators to encode relative positioning of signs, as found in the Manuel de Codage (see §2 below for further discussion of the Manuel de Codage). With proposal N1944 (Everson 1999a), this was extended to several thousands of signs, incorporating signs from Grimal *et al.* (1993). Both proposals drew quite some criticism, for example from Wolfgang Schenkel (Schenkel 1999; Everson 1999b).

The third attempt was more modest, with a total of 1071 signs, including the signs from Gardiner (^{3}1957) plus those from its supplements (Gardiner 1928, 1929, 1931, 1953) and a few signs from other sources, but leaving out the formatting operators. This list was finally approved and added to Unicode

[*] I gratefully acknowledge fruitful discussions with Serge Rosmorduc. I am also very grateful to Horst Beinlich and Norbert Stief for correspondence about PLOTTEXT. Many thanks go to anonymous reviewers for a large number of improvements to the text.

5.2 as part of Plane 1 (Unicode 2010). (Plane 1 is a range of code points that is mostly used for characters from historic scripts.)

One factor that marred the discussions leading up to the eventual Unicode set of hieroglyphs was the disparity between the formal notion of 'character' and standard practices in Egyptology when transcribing hieratic or normalising hieroglyphic inscriptions. Following the terminology of Unicode, a *character* is the smallest component of written language and a *glyph* is a shape that a character can have when it is rendered or displayed. In Egyptology however, there seem to be tendencies to remain true to the original manuscript while encoding a text, often to the extent of encoding glyphs rather than characters.

An example is the distinction between 𓅱 (G43) and 𓏲 (Z7), which could be argued to be different shapes representing the same character. One also sees occurrences of 𓏃 (W17) next to 𓏅 (W18), which are different glyphs for the same character.

The cause of much of the confusion is the sign list by Gardiner, or perhaps more accurately put, its current misinterpretation. The intention of this list was never to create a list of characters in the sense of Unicode, but firstly to offer students an overview of different hieroglyphs and their functions and meanings, and secondly, to create an inventory of signs needed to print texts. In particular, for some signs, there is more than one glyph to be used by the printer in different contexts, such as 𓅩 (G36a) next to 𓅨 (G36) and 𓈋 (N25a) next to 𓈊 (N25).

The problem of what information to represent in an encoding of hieroglyphic text not only pertains to the sign list but also involves the formatting, or in other words, how signs are positioned relative to one another. It is not always clear to what extent this aspect is important to encoding: on the one hand the relative positions of signs have little linguistic significance, whereas on the other hand it is standard practice to remain true to the formatting of the original text. Rare examples when relative positioning does have linguistic meaning include 𓏌𓐰𓈗.

A further complication is that different intended applications call for different levels of information to be present in hieroglyphic encodings. Examples of applications include:

- Studies in palaeography and epigraphy.
- Study of the translation of a particular text.
- Study of grammar.
- Lexicography.

For palaeography and epigraphy, one would wish to preserve as much as possible of the physical appearance of signs as well as their relative positioning.

For translations, a normalised hieroglyphic rendering is usually sufficient. Where there is doubt about its accuracy, one may wish to compare it to a facsimile of the original manuscript. It is easy to find the relevant fragment of the facsimile on the basis of the normalised rendering provided the latter preserves an appropriate amount of the formatting of the original.

For the study of grammar, much of the appearance of hieroglyphs and their relative positioning are of little relevance. Nevertheless one wishes examples in a grammar book to conform generally to conventions of hieroglyphic composition in order to give an accurate impression of the written language.

In lexicography, the attempt is usually made to abstract away from the formatting of particular instances of words. Ancient Egyptian lacks a notion of orthography in the sense of having a single correct writing, and one word may be written with different sequences of hieroglyphs, even within a single text. Consequently, a lemma in a lexicon may consist of an idealised hieroglyphic writing, possibly without any formatting information at all.

The above four example applications illustrate different sets of requirements one may want to impose on an encoding scheme for hieroglyphic text, some with an emphasis on faithfulness to one particular manuscript, others with an emphasis on uniformity across manuscripts.

Other aspects of this discussion include the versatility of encoding schemes and, related to this, the lifespan of encodings. For example, a representation of an hieroglyphic text that is close to a facsimile, with precisely specified scalings and positionings of signs and custom drawings of non-standard glyphs is not very suitable for applications of automatic processing, such as compilation of word lists, automatic transliteration, etc. Such 'pseudo-facsimile' representations also tend to heavily rely on one particular choice of font, and often on one particular software tool offering certain functionality to indicate relative positioning of signs. This severely limits the lifespan of the encodings, as tools and fonts are typically replaced by others after a relatively short time.

However, it is not self-evident that the lifespan of pseudo-facsimile encodings is an issue in practice. In a typical scenario, one could compile a faithful encoding of a manuscript, then convert this to a general-purpose graphical format, such as JPEG or PDF. This can be included in a publication of the manuscript. Thereafter one may safely discard the encoding as it has few other uses.

In this article we will consider encoding from an entirely different perspective, namely that of creating and maintaining a corpus of hieroglyphic texts that has a reasonable life expectancy and can be used for various applications. These applications are numerous: not only the publication of the texts themselves, in electronic format or on paper, but also the reuse of the material in learning and teaching, extraction of sentences for the use in grammar books, extraction of words for use in lexicography, etc.

Some requirements for such an encoding scheme with both longevity and versatility are:

- stability,
- high expressive power,
- font-independence,
- simplicity,
- precision of meaning, and
- flexible formatting.

The need for an encoding scheme that is stable is obvious. In a large corpus that is under development, it would be impractical if frequent modifications to the corpus were required as a result of changes to the encoding scheme. Connected to this is the need to make the encoding scheme powerful enough to deal with most if not all texts that one may reasonably expect to encounter.

Due to the open-ended nature of hieroglyphs, there is no hope of compiling a 'complete' sign list. However, one would expect the expressive power of the encoding scheme to at least cover most if not all kinds of relative positioning that one finds in practice.

A hieroglyphic font is generally a stylised idealisation of the signs that can be found in good monumental inscriptions. Due to the large diversity of styles across periods and regions, it is unlikely that one font will ever satisfy all scholars. Furthermore, a detailed font with fine lines may be more suitable for printing on paper whereas a less detailed font with thick lines may lend itself better to use on computer monitors. In order to use an encoding in a wide range of applications, it should therefore be independent of a particular font.

Data tends to outlive the software by which it is created. Often this is because programming languages can become obsolete very quickly. It is therefore necessary to use simple data formats for which new processing software can be developed easily. The correctness of this software can be guaranteed if the meaning of constructions in the data formats is precisely defined.

Lastly, some applications, such as alignment with transliterations, require provisions for automatically inserting whitespace within hieroglyphic encodings. However, the encoding scheme itself should be free of physical linebreaks and pagebreaks, leaving it to each application to determine appropriate places for these.

As illustrated by examples in the following sections, the issue of the sign list cannot be seen as independent from the issue of formatting, at least in many existing encoding schemes. In many cases, inadequacies in operators for relative positioning have led to addition of spurious variant glyphs or combinations of signs. In addition, applications ranging from pseudo-facsimile reproduction to lexicographical analysis are also relevant to these encoding questions: not only which signs (characters or glyphs) should be included, but also what kinds of relative positioning need to be available.

2. WHY THE MANUEL DE CODAGE IS INADEQUATE

It is difficult to talk about a single Manuel de Codage (MdC) encoding of hieroglyphic. This is because the last published version was from 1988 (Buurman *et al.* 1988), henceforth referred to as MdC88. Since then many features have been added to hieroglyphic editors but without proper documentation. Some of these editors were developed by the CCER. One phrase on page 15 of Buurman *et al.* (1988) is particularly revealing:

> [...] the Glyph programme [sic], linked to this enterprise from the beginning, has been improved, which had to be included in the Manual

This suggests that the MdC was not intended as a standard in itself, but rather as a manual for a particular tool. In addition, there are by now many competing hieroglyphic editors, each adding its own features and interpreting various imprecisely documented features from MdC88 in different ways.

Rather than directly criticising the MdC or any of its dialects, it is perhaps more appropriate to criticise the tradition of hieroglyphic encoding starting with Buurman *et al.* (1988). The most serious defects within this tradition are:

- The encoding schemes are specific to particular versions of particular tools.
- The emphasis is on creating pseudo-facsimiles. Long-term storage of hieroglyphic encodings for diverse usage and for reuse has low priority.
- Connected to this, the font used is the one that came with the tool. Exchanging one font with another is not guaranteed to give a satisfactory appearance.

A case in point is the operator `&`. It is not part of MdC88, but it has been part of implementations of Glyph for a long time. It occurs in the expression `G14&X1` in an unfinished, updated Manuel de Codage by Hans van den Berg (1997). The operator can be used to separate two or more occurrences of hieroglyphs. Its meaning is undefined except for a finite set of sequences of hieroglyphs specific to the hieroglyphic editor. Where this meaning is defined, it is a particular relative positioning and/or scaling of the individual hieroglyphs. It is typically used where the two operators `:` for vertical and `*` for horizontal combination do not suffice.

The problem is that the number of combinations of glyphs for which the `&` is needed is potentially unbounded. To put it in another way, if we define an expression with `&` for every occurrence of a hieroglyphic group that cannot be described as purely horizontal or purely vertical arrangement of subgroups, then encoding any new text will require defining new expressions. This makes the encoding scheme unstable to the extreme.

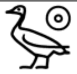
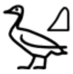
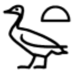
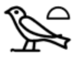
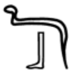
group	EGPZ	RES
	G39&N5	insert[te](G39,N5)
	G39&N29	insert[te](G39,N29)
	G39&X1	insert[te](G39,X1)
	G36&X1	insert[te](G36,X1)
	I10&D58	insert[b](I10,D58)

Table 1. Groups that are not formed by purely horizontal or vertical arrangements, their expressions in the EGPZ, and their expressions in RES (see §3)

Tab. 1 shows a few examples of expressions with & out of the no less than 400 such expressions included in the EGPZ (Saqqara Technology 2008). This is of course nowhere near an exhaustive list of combinations of glyphs for which the operators : and * do not suffice. The problem is the lack of power of the latter two operators, in combination with a possible misconception that horizontal and vertical relative positioning would be the norm in hieroglyphic writing, and other types of relative positioning would be the exception. Even a cursory glance at a few original hieroglyphic inscriptions will immediately refute this assumption, as the so called 'special' groups are very common.

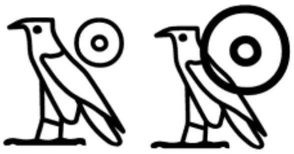


Table 2. The risk of hard coding of scaling factors and absolute positions. What may look satisfactory with one font (left) may be entirely unsatisfactory with a different font (right)

Some dialects of the MdC have tried to solve this problem with hard coding of a scaling factor and an absolute position for each occurrence of a hieroglyph in a 'special' group. The problem with this is that the life expectancy of such an encoding does not extend beyond the lifespan of the font with which the choice of scaling factors and positions were determined. This is illustrated in Tab. 2, assuming two different fonts in which the sun-symbol has different sizes.

The Manuel de Codage has more shortcomings, such as the lack of standardisation and the cumbersome syntax, which make it difficult to develop parsers and renderers. It is also problematic that the Manuel de Codage was designed as a holistic file format, to be used for document preparation, including operators for hard linebreaks and pagebreaks. Had the MdC been restricted to just hieroglyphic encoding to be used within arbitrary file formats, it would have inspired more flexible usage, for example for automatic analysis and lexicography.

Some of these shortcomings can be fixed to a certain extent. For example, one could imagine that the Egyptological community as a whole would at some point agree on a common standardised dialect of the Manuel de Codage. However, the traditional emphasis on pseudo-facsimiles and the assumption that encodings are discarded after publication of a text have had too great an influence on the development of common MdC dialects. A substantial paradigm shift is needed to arrive at an encoding scheme that offers any hope that text encodings might survive a change of font or a change of hieroglyphic rendering tool.

3. PROPOSED SOLUTION

The Revised Encoding Scheme (RES) was introduced in Nederhof (2002) and criticism on it was addressed in Nederhof (2008). The development took place in three stages.

First, we investigated large amounts of hieroglyphic texts, as well as modern (hand-drawn) transcriptions of hieroglyphic and hieratic texts. The purpose of the latter was to find out which aspects of formatting of hieroglyphic texts Egyptologists typically want to preserve. We have deliberately ignored typeset hieroglyphic texts, as those are commonly fettered by technological limitations of the formatting and printing tools that were used.

In the second step we designed a small set of operators to express relative positioning of hieroglyphs, such that in principle all of the 'special' groups we found in real texts can be expressed using a combination of those operators. This has been done without too much concern for the technical difficulty of the implementation of the operators.

The technical realisation came in a third step. Whereas the implementation of the most innovative operators can be difficult, it should be pointed out that this task needs to be done only once, and is outweighed by the ease with which texts can be encoded and the ensuing longer lifespan of encodings, independent of any font.

The font-independence comes from the design decision that the meaning of operators should match observable arrangements of signs. For example, one use of the `insert` operator corresponds to the intuitive arrangement that can be described as 'one sign is to be placed in the free upper-right corner next to another, and scaled appropriately'. Encoding can thereby be done by visual inspection rather than by dragging images by the mouse. The consequence is that the unfortunate situation in Table 2 is avoided.

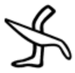
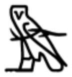
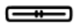
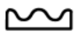
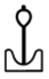
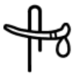
group	Unicodc	RES
	D57	stack[on](D56, T30)
	M3a	stack(G17,M3)
	N18b	inscrt[scp=0.0](N18,O34)
	N25a	N25[yscalc=0.8]
	T3a	T3 :[scp=0.0,fit,fix] N26
	T33a	stack[under](T33, S29)

Table 3. Groups that have been given their own code points in Unicode 5.2, but that can be described equally well by RES expressions

Tab. 1 already presented examples of the use of the `insert` operator. Tab. 3 presents further examples of groups of signs that can be expressed in terms of combinations of more elementary signs using RES operators. These groups have in fact been given explicit code points in Unicode 5.2. By our reckoning, there are 105 such groups out of the 1071 hieroglyphic code points in Unicode 5.2 (Nederhof 2011). This strongly suggests that future extensions of the sign list can remain much more modest and manageable if an encoding scheme such as RES is adopted in place of MdC. It should further be pointed out that overly large sign lists with large portions of extraneous signs and sign combinations, such as the EGPZ mentioned in §2, place an unreasonable and unnecessary burden on font developers.

There are provisions in RES for fine-tuning aspects of the formatting, such as an indication that the distance between two signs should be, say, half or double what it would normally be. This can be used for pseudo-facsimile representations, which may be ill-advised for all but a few applications. One may deliberately want to avoid this type of fine-tuning for most applications. If such fine-tuning is used, it will under normal circumstances not be invalidated by a change of font in the sense that a 'wrong' rendering as in Tab. 2 would be produced.[1]

Lastly, it should be pointed out that great advancements towards more powerful hieroglyphic encoding schemes were already made in PLOTTEXT (Stief 1985). In that system there are, for example, operators for placing a sign in a free corner next to another sign, comparable to our `insert` operator.

4. DISCUSSION

The creation of large electronic corpora of hieroglyphic texts is only cost-effective if the validity of the encodings can be preserved over a long period. In the tradition of the Manuel de Codage, the validity of an encoding is specific to a certain choice of software package and font, which precludes longevity of the electronic resources. Consequently, if there is to be any hope of developing comprehensive corpora, the Egyptological community should abandon the Manuel de Codage encoding of hieroglyphic text. One viable alternative in the form of RES is readily available.

There are currently no well-defined criteria by which one can decide which new hieroglyphs should be added to the Unicode set. Developing such criteria is all the more difficult as the character/glyph dichotomy seems to be far apart from the way that hieroglyphic texts are commonly transcribed, for most relevant applications. It is also possible that systematic investigations of shapes and meanings of signs, such as those by Meeks (2004), will one day bring us closer to an answer. What does seem clear is that a well-designed encoding scheme will avoid the need for extraneous signs, added just to compensate for the inadequacies of the relative positioning operators.

BIBLIOGRAPHY

VAN DEN BERG, Hans. 1997. Manuel de Codage: A standard system for the computer-encoding of Egyptian transliteration and hieroglyphic texts, in: http://www.catchpenny.org/codage/ (accessed 2011-09-30).

BUURMAN, Jan, Nicolas GRIMAL, Michael HAINSWORTH, Jochem HALLOF & Dirk VAN DER PLAS. 1988. *Manuel de codage des textes hiéroglyphiques en vue de leur saisie sur ordinateur*, Paris, Institut de France.

DANIELS, Peter T. & William BRIGHT (eds.). 1996. *The World's Writing Systems*, New York, Oxford University Press.

EVERSON, Michael. 1997. Proposal to encode basic Egyptian hieroglyphs in Plane 1, in: ftp://std.dkuug.dk/jtc1/sc2/WG2/docs/n1637/n1637.htm (accessed 2011-09-30).

—. 1999a. Encoding Egyptian hieroglyphs in Plane 1 of the UCS, in: http://std.dkuug.dk/jtc1/sc2/wg2/docs/n1944.pdf (accessed 2011-09-30).

—. 1999b. Response to comments on the question of encoding Egyptian hieroglyphs in the UCS (N2096), in: http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2132.htm (accessed 2011-09-30).

GARDINER, Alan H. 1928. *Catalogue of the Egyptian Hieroglyphic Printing Type, From Matrices Owned and Controlled by Dr. Alan H. Gardiner*, Oxford, Oxford University Press.

—. 1929. Additions to the new hieroglyphic fount (1928), in: *The Journal of Egyptian Archaeology* 15, p. 95.

—. 1931. Additions to the new hieroglyphic fount (1931), in: *The Journal of Egyptian Archaeology* 17, p. 245-247.

—. 1953. *Supplement to the Catalogue of the Egyptian Hieroglyphic Printing Type, Showing Acquisitions to December 1953*, Oxford, Oxford University Press.

—. ^{3}1957. *Egyptian Grammar: Being an Introduction to the Study of Hieroglyphs*, Oxford, Griffith Institute.

[1] More information on RES can be found at: http://www.cs.st-andrews.ac.uk/~mjn/egyptian/res/.

GRIMAL, Nicolas, Jochen HALLOF & Dirk VAN DER PLAS. 1993. *Hieroglyphica*, Utrecht-Paris, Publications Interuniversitaires de Recherches Égyptologiques Informatisées.

MEEKS, Dimitri. 2004. *Les architraves du temple d'Esna. Paléographie*, Cairo, Institut français d'archéologie orientale (= Paléographie hiéroglyphique 1).

NEDERHOF, Mark-Jan. 2002. A revised encoding scheme for hieroglyphic, in: *Proceedings of the 14th Table Ronde Informatique et Égyptologie*. On CD-ROM.

—. 2008. Automatic alignment of hieroglyphs and transliteration, 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*, Gorgias Press, p. 71-92.

—. 2011. The 1071 hieroglyphs from Unicode 5.2, in: http://www.cs.st-andrews.ac.uk/~mjn/egyptian/ unicode/ (accessed 2011-09-30).

SAQQARA TECHNOLOGY. 2008. EGPZ 1.0 specifications, in: http://www.egpz.com/resources/egpz.htm (accessed 2011-09-30).

SCHENKEL, Wolfgang. 1999. Comments on the question of encoding Egyptian hieroglyphs in the UCS, in: http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2096.pdf (accessed 2011-09-30).

STIEF, Norbert. 1985. Hieroglyphen, Koptisch, Umschrift, u.a. – Ein Textausgabesystem –, in: *Göttinger Miszellen* 86, p. 37-44.

UNICODE CONSORTIUM. 2010. Egyptian hieroglyphs, in http://www.unicode.org/charts/PDF/U13000.pdf (accessed 2011-09-30).

Hieroglyphic Text Corpus

Towards Standardization

Vincent EUVERTE & Christian ROY

Rosette Project – Paris

1. INTRODUCTION

The I&E meeting in Liège was the eighteenth of a long series. Looking back to these 26 years since the first meeting of this group in the College de France in Paris (1984), three great periods can be recognised:

(1) Two events actually happened 15 years earlier, in 1969, with the appearance of Arpanet (the ancestor of the Internet) and of the first Glyph program by Jan Buurman on a mainframe in Algol language. In the late seventies, information technology has been moving from the domain of industry to that of personal computing. So Jan Buurman migrated his program to Fortran, in order to be transferrable to PCs. The I&E Computer group then contributed at that time by launching the Manuel de Codage, the 3rd edition of which was presented in Cairo at the 1988 Congress of the International Association of Egyptologists. At the same time, the first electronic dictionary of Ancient Egyptian was launched with the *Wörterbuch*.

(2) A second period led this group to Bordeaux, for its 10th anniversary in 1994. Among the major achievements, we will cite the creation of the Multilingual Egyptological Thesaurus (MET, see §2) which was supplemented by a list of 14 "minimum requirements" named "passeport".[1] The publication of the "Beinlich wordlist"[2] was also a significant outcome of this meeting. In his retrospective of these ten years, Dirk van der Plas also raised the idea of a fourth edition of the Manuel du Codage, which was never realised. As technology was progressing quite fast and Internet usage was rocketing up to 10 million users, several hieroglyphic text processors appeared in parallel. We cannot cite them all, but the best known are WinGlyph and MacScribe, or Hierotext migrating to TKSesh (see Gozolli's overview in the present volume).

(3) The third major period is less clear to us as we have not been able to retrieve all the proceedings from these 15 years. Major outcomes seem to be the very long debate on including hieroglyphs in Unicode, finally endorsed in the Fall 2009, thanks to the persistence of Michael Everson and Bob Richmond. During this period, several online dictionaries also appeared, and large developments in computer technology encouraged the appearance of new software, as JSesh for example. Various communication tools, such as AEL, EEF, ThotScribe fora and Internet letters, as well as the multiplicity of databases on mastabas, shawabtis, pyramid texts, etc. (including the *Ramsès* Project) can be added to figure 1.

[1] In: *Actes des Rencontres "Informatique et Égyptologie" 1993*, Informatique et Égyptologie 9, 1994, p. 4.

[2] Beinlich, Horst & Friedhelm Hofmann. 1994. Ägyptische Wortliste, in: *Göttinger Miszellen* 140, p. 101-103.

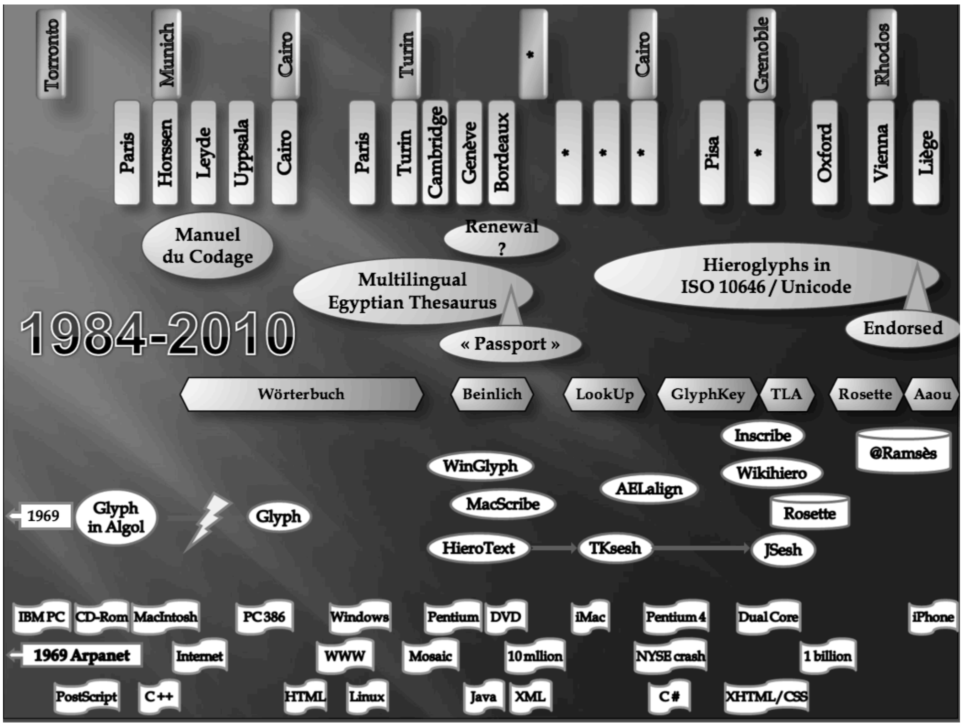


Figure 1. Informatique & Égyptologie timeline

A quick look at figure 1 makes it obvious that sharing the heritage of Ancient Egyptian written production is facing numerous technical challenges in a constantly evolving environment. We will focus in this paper on two specific topics related to the standardization of Ancient Egyptian text corpora: the 'Multilingual Egyptological Thesaurus' (MET) and the 'Manuel de Codage' (MdC).[3]

2. THE 'MULTILINGUAL EGYPTOLOGICAL THESAURUS' (MET)

2.1. *What is an ancient Egyptian object?*

In our field, the "Objects" of study are artefacts with many facets seen from epigraphic and philological viewpoints. To characterize such objects, it is essential to describe their contexts and environments: Where were they discovered? Where are they now? Who found them? What are they made of? To which period do they belong? In addition, things like pictures of the artefact and associated bibliography are necessary to enrich the description.[4]

An excellent base for communality of the discipline was defined in the nineties with the Multilingual Egyptian Thesaurus[5] (MET).

According to the MET definition, an 'object' is any Museum artefact. In a broad sense, it includes objects such as the White Chapel of Senousret in the open air museum at Karnak. It also encompasses any papyrus from any period, even if in a private collection. But what about the objects which are still sitting on their original site, such as an Amarna Boundary Stela, the Famine Stela on Sehel island, or (a bit more complicated) the Qasr Ibrim stela relocated to the new site of Kalabsha? We may also want to characterize similarly the context of reliefs and paintings in temples and tombs or even petroglyphs

[3] We would like to highlight that both projects originated many years ago from groups represented at this conference, and both are still in use in various capacities, regardless of weaknesses, obsolescence, or hiatus.

[4] More specifically for a text corpus, the content includes several categories specific to the hieroglyphic writing: the characters must be identified, their graphic representation may vary according to the period, the support, the type of text, and the layout may even include scenes which are an integral part of the meaning. These are all aspects addressed by the Manuel de Codage (see section 3 below).

[5] http://www.ccer.nl/apps/thesaurus/index.html.

located everywhere in the Egyptian deserts: boats and serekhs in Wadi el Shott, ostriches in Aswan quarry, or quarryman's marks in Gebel Silsileh.

2.2. *MET usage: Who?*

Who are the current and potential users of the MET? Of course museums at first, as the MET was created for them; but from previous examples, one can imagine dozens of other applications of this method of cataloguing.

Who may contribute, propose amendments or new data? To our mind, anybody interested in Egyptology has this potential. However these contributors are not necessarily legitimately authorised to endorse and to publish. So we see three major steps in the Thesaurus management:

- *Collection*. As suggested by Reem Baghat, responsible for the MET at CULTNAT,[6] a tool could be implemented, for instance on the GEM (Global Egyptian Museum) website,[7] to allow identified/registered users to record their information/proposals.
- *Validation*. A committee of professionals is definitely required, and it must be international to ensure proper perspectives and translations in each of the agreed languages.
- *Distribution*. The most up-to-date approved version of the Thesaurus could be made publicly available in an exchangeable format (e.g. PDF). The GEM website should then probably be the most appropriate medium.

2.3. *MET completion*

As we saw with the timeline (see §1), no official update has been made since the 1995 publication, which raises several issues:

- The "Provenance" characteristic is missing quite a number of locations, either neglected initially such as Lower Nubian sites now under the Nasser Lake or recently excavated as Tell-Herr in Sinai.
- For the "Current Location" field, several new museums have opened in the past 15 years, such as the Imhotep site museum in Saqqara. Some others have been forgotten in the initial list such as the small Tessé museum in Le Mans (France).
- The Thesaurus details precisely the different types of support material; but how to indicate a David Robert's painting or a 19th century facsimile describing an object no longer available to us, because it has been eroded, robbed or destroyed?
- The philologists could make suggestions in order to expand the current "Language" and "Writing" characteristics of the Thesaurus, for instance to allow the description of the "state of the language" depending on the period and the type of text.
- Last but not least, among the 7 languages already defined, some translations are either incomplete or inconsistent, in particular in Portuguese.

2.4. *MET expansion*

There may also be a lot of enrichment to the current 15 dimensions of the Thesaurus:

- For the sake of clarity, the existing characteristics could be enriched, for instance with dating criteria and with the Ancient Egyptian and Greek names added to the Arabic names (as much as we know them).

[6] Center for Documentation of Cultural & Natural Heritage; see http://www.cultnat.org/.

[7] http://www.globalegyptianmuseum.org/.

- The Global Egyptian Museum made a tremendous effort to expand the original MET with interesting characteristics, such as Colour, Culture, Titles, Dimensions, etc. (some of them been integrated in the "Passeport" definition of 1993; see n. 2 above).
- The Global Egyptian Museum also launched the translation of the Thesaurus in Arabic. One can even envisage adding other languages, like Japanese and Chinese, as these countries are becoming more and more involved in archaeology of Egypt.
- Regarding the bibliographic references, the I&E group is already supporting the AEB/OEB as the most international standard.

2.5. *MET Revival Project: Who and When?*

- Since ownership of the MET was transferred from the CCER to CULTNAT, it is clearly the new leader and also has the authority and resources for the web deployment of a MET revival project using the Global Egyptian Museum Internet facilities.
- The I&E Computer Group could serve as a facilitator to identify professionals in each country to complete/validate/translate. A clear message from us, eventually supported by the IAE and its President James P. Allen, could convince CULTNAT to go ahead.
- With an open but controlled Internet interface, many professional contributors may simplify the collection task; one may even envisage the contribution of benevolent amateurs with regard to data collection, pending a final review by a validation team composed of international professionals. Such an approach may significantly reduce the necessary budget for this project.

3. MANUEL DE CODAGE (MDC)

Looking back at the third edition of the MdC, published in 1988, it appears that there are essentially three parts in this document, and we are not convinced that they all belong to the same matter:

- The phonetic values should be part of the user-interface, so something to be managed at the software level, rather than concerning the language itself.
- The sign list is perhaps an endless debate and is not within the scope of this paper. Let's just notice that the acceptance by Unicode of a basic list of 1100 signs is already a significant step on the long path to communality.
- The third part refers to the MdC coding aspect: how to identify a sign and to represent it in the appropriate position. This will be the main focus of the present discussion.

3.1. *The MdC coding aspect*

This part of the paper focuses on the 'syntax' and associated 'semantics' used to code hieroglyphic texts with the aim to display or print them. The text coding is entered by the users:

- either through specialised graphical interfaces
- or through standard text editors.

This results in an 'External viewable coding',[8] where:

- 'External' means 'easily exchangeable between computers and between software';
- 'Viewable' means that 'it can be directly read and understood by human users'.

[8] Some rendering engines may choose to use an internal coding. Here we will not consider any internal coding which constitutes specific implementation details.

Our subject here is the syntax and the semantics used in this ‘External viewable coding’. So far this coding is based on a formal description elaborated in 1988 and named ‘Manuel de Codage 88’.[9]

3.2. *Basic Requirements*

MdC requirements must be considered together from a user’s viewpoint and from the rendering engine’s perspective:

From a user’s view point:

- coding must be easily understood and learned: a modern approach allows for a simpler syntax than the original MdC88;
- syntax must not be too verbose to input easily into standard text editors;
- two sets of functionalities should be distinguished:
  - *basic* functionalities are fulfilled by any rendering software with a standardised syntax;
  - *extended* functionalities are not mandatory but, when supported, are based on a standardised syntax;
- the current MdC88 syntax must be supported for a reasonable period of time.

From a rendering engine’s perspective:

- a ‘regular’ syntax affords the above benefits and makes it possible to produce efficient software using fewer system resources, and can be more economically developed and maintained (based on standard tools);[10]
- software does not display errors when an unsupported extended functionality is met;
- software should provide a tool analysing the code and diagnosing ‘deprecated’ (see below), i.e. unsupported, functionalities and syntax errors.

3.3. *Basic and Extended Functionalities*

Functionalities may be categorized on an axis ‘basic/extended’. Below are several examples of functionalities:

- fragments vs facsimile: the simpler (more basic) functionality in this respect is to render fragments of hieroglyphic texts one by one without assembling them like in a facsimile (more extended) where fragments are combined with their relative positions, orientations, directions of writing, etc.
- simple vs complex cadrats: a complex cadrat requires a precise control of the size and positioning of inner subcadrats.
- simple vs complex alignments: a simple alignment is setting a position relative to the current position when the alignment is already specified (rather similar to word processing tabulation). A complex alignment takes into account actual size of the components being aligned.
- no text vs integrated texts — we refer here to texts for comments, transliteration, etc.: as an extended functionality, such text may be embedded within hieroglyphic texts.

[9] Many extensions without effective standardization have been made to the MdC88 by various software programs.

[10] Precise definition of a ‘regular’ syntax is out of the scope of this paper (see http://en.wikipedia.org/wiki/Backus-Naur_Form).

3.4. *Developments*

To describe MdC developments, we will cover the two following 'directions' of development:

- new functionalities (§3.4.1),
- new updated syntax (§3.4.2).

3.4.1. New Functionalities

We will give three examples of potential new functionalities:

- Vertical alignments. In some circumstances, it may be very useful to vertically align hieroglyphs and transliteration, for example.[11]

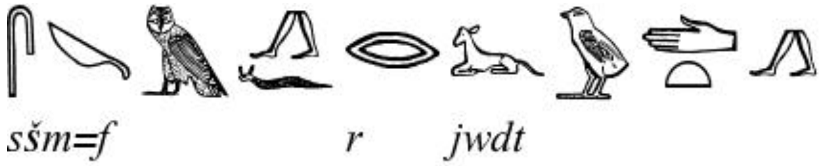


Figure 2. Example of vertical alignments

- Horizontal alignments. Fig. 3 displays a facsimile of the top part of the south face of the Luxor obelisk in the Place de la Concorde (Paris). It is composed of one fragment with three columns: horizontal alignments of the corresponding sections (for example cartouches) are very likely a new functionality desirable for the new MdC.

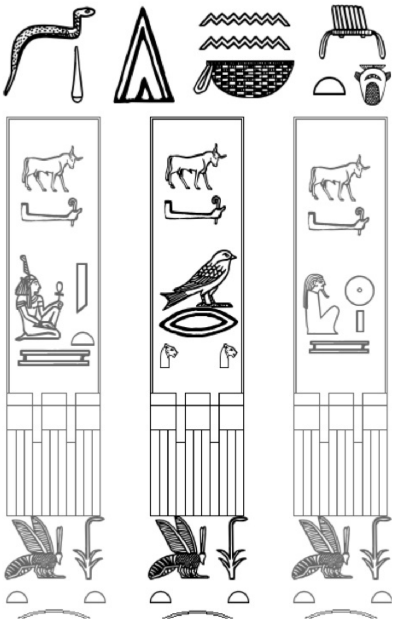


Figure 3. Example of horizontal alignments

- Browsable facsimile: several text fragments with different size, orientation, etc. are combined like in the original artefact. In addition the facsimile is browsable.

[11] See Nederhof, Mark-Jan. 2009. Automatic creation of interlinear text for philological purposes, in: *Traitement automatique des Langues* 30/2, 237-255.

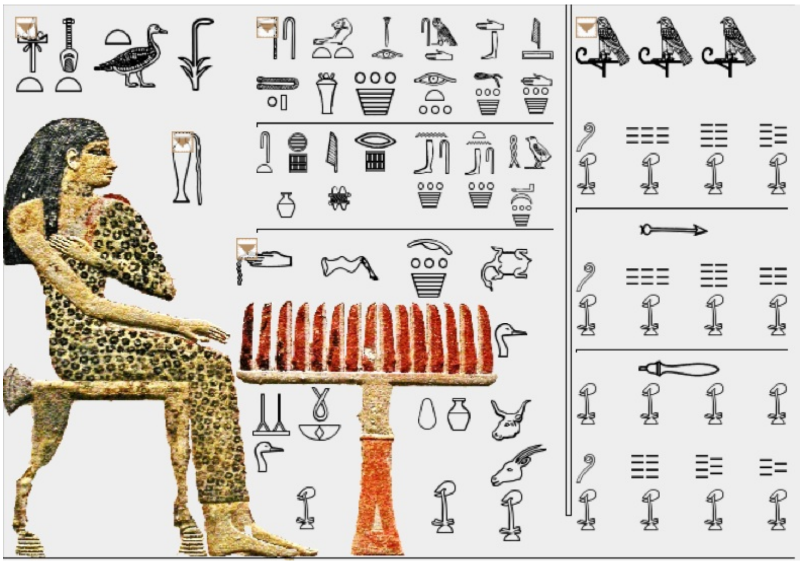


Figure 4. Facsimile of the Nefertiabet Stela

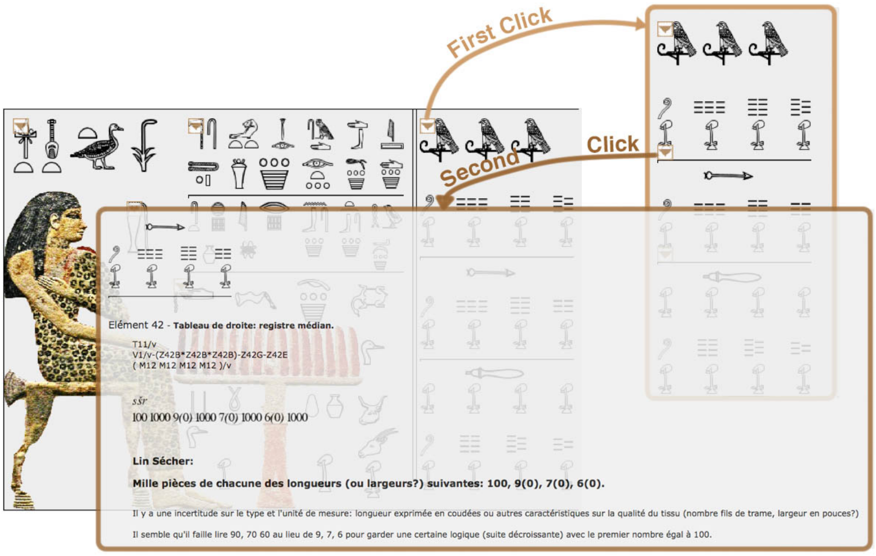
When initially displayed, brown rectangles are drawn to delimitate ‘elements’ of the stela. For instance, to the right of the offering table are three horizontal sub-elements. On the left top corner of each rectangle is drawn a brown ‘down’ arrow which is clickable to explore the corresponding element. Browsing the table on the right side of the stela follows three steps:

Step 1: the full stela is displayed → click on the down arrow of the right-most table.

Step 2: the right most table is displayed. Note that the picture of this table is shown to the right (Hieroglyphic colouring could be an extended functionality) → click on the down arrow of the middle register.

Step 3: the middle register is displayed with transliteration, translation, comments, etc.

At steps 2 and 3, an ‘up’ arrow is displayed to return to the previous step. These three levels are defined in the text/facsimile MdC coding. The arrows used for navigation are interface elements not in the scope of the MdC coding.[12]



Elément 42 - **Tableau de droite: registre médian.**

T11/v
V1/v-(Z42B*Z42B*Z42B)-Z42G-Z42E
(M12 M12 M12 M12)/v

sšr
100 1000 9(0) 1000 7(0) 1000 6(0) 1000

Lin Sécher:

Mille pièces de chacune des longueurs (ou largeurs?) suivantes: 100, 9(0), 7(0), 6(0).

Il y a une incertitude sur le type et l'unité de mesure: longueur exprimée en coudées ou autres caractéristiques sur la qualité du tissu (nombre fils de trame, largeur en pouces?)

Il semble qu'il faille lire 90, 70 60 au lieu de 9, 7, 6 pour garder une certaine logique (suite décroissante) avec le premier nombre égal à 100.

Figure 5. Browsing in a facsimile

[12] This static printed presentation cannot demonstrate the whole process. A live example is available at: http://projetrosette.info/page.php?Id=799&TextId=134&line=1&nbrElts=1.

Quite complex facsimiles may also be produced using this technique, as in the example of the astronomic ceiling of the Ramesseum in fig. 6.

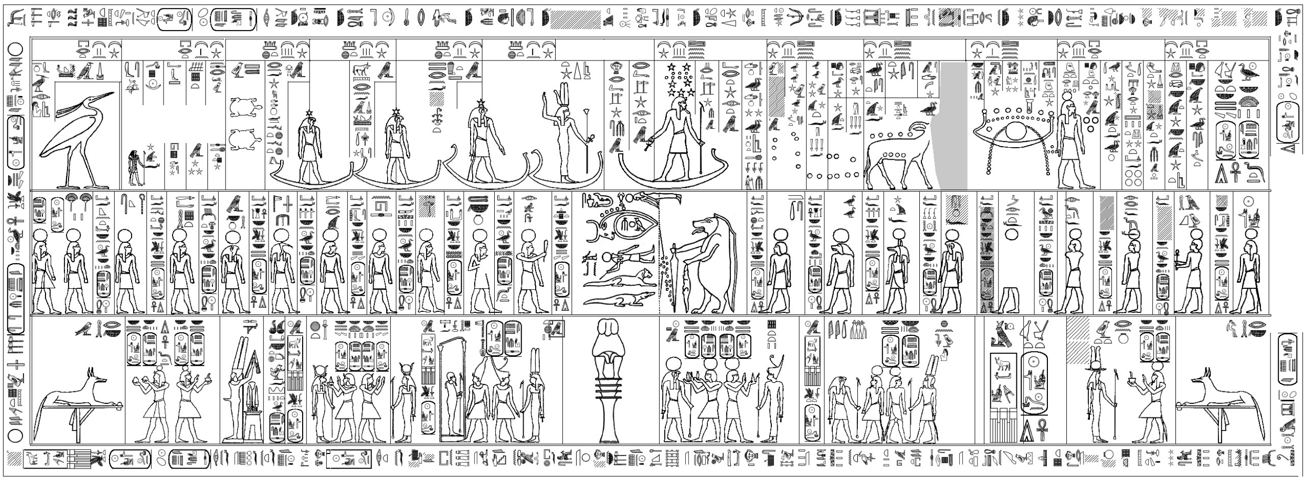


Figure 6. Astronomic ceiling of the Ramesseum

Facsimiles can be exported to external files with high definition and common graphic formats for inclusion in printed documents (facsimile of the fig. 6 has been printed on a poster 2.5 meters wide).

3.4.2. Move to an updated syntax

The basis of MdC88 is combining:

- symbols for hieroglyphs (Gardiner codes, phonetic equivalents)
- operators for positioning (‘-’, ‘ ’, ‘*’, ‘:’, …)

This complies with all our ‘basic requirements’ described above. For ‘modification’ of hieroglyphs, however, MdC88 uses a ‘chaotic’ syntax, unable to fulfil these same requirements. For instance:

- the same character may signify totally different functions (e.g. ‘#’ is used for both superposition and hashing);
- any new functionality requires the choice of a new ‘character’ in a more and more limited set. For instance, to colour a hidden or partially erased glyph in grey we use `$g`, which will limit further possible colours, is ambiguous with green, and does not allow us to cover the full colour space;
- this ‘irregular’ syntax requires more system resources and practically prevents the use of standard tools for the MdC syntax analysis (see http://fr.wikipedia.org/wiki/Lex_et_yacc).

It makes then sense to move to an updated syntax. We suggest four guidelines for evolution to such syntax:

- keep the ‘foundation’ syntax (described above): an important part of the existing coding remains valid;
- mark inappropriate syntax elements as ‘deprecated’: MdC88 syntax elements like ‘#’, ‘$b’, ‘$r’, and many others may still be used but for a limited period of transition;
- implement a new and consistent syntax;
- fully support MdC88 during the same transition period.

3.5. *Implementation of these principles in the Rosette Project*

We will now give additional information about the implementation of the above principles on the Rosette Project web site.[13] A few new elements are added:

- 'modification' operators allow us to modify the rendering of affected hieroglyphs. Inserted just after the modified hieroglyph, they combine a '/' with a letter determining the modification:
  - 'c' for colour
  - 'r' for rotation
  - 'a' for hashing
  - etc.

  The letter is followed by relevant parameters like `/cr` for colour red or `/c255,0,0` for an RGB colour. These operators may be 'factorised' to several hieroglyphs enclosed between parentheses, for example: `( A1 A2 A3 )/cr/r45`: the three hieroglyphs will be drawn in red and rotated by 45°.
- '# tags' modify the drawing state. For instance `#poV;x=5;y=5 A1 A2 A3` draws A1 A2 A3 vertically starting from x=5 and y=5. A few other # tags are available. For example `#ssr;x2;y2;h3;w4` will draw a rectangle from x=2, y=2 with an height of 3 and a width of 4.[14]
- 'texts' may be mixed with hieroglyphs and can be amended by modification of operators and # tags.

The two following characteristics should also be mentioned:

- A syntax checker signaling deprecated elements is available.
- Integrated support of Unicode 5.2.

As an illustration, we will show the coding used for two examples:

(1) Stela of king Kamose: on top of this stela, an 'ankh' sign has been overwritten. To render this overwriting, we will use the following code: `M4 (t:3)*anx/y25/s50/x18/c100 G5 xa:a Hr:1 g:f`. Four modification operators are applied to the ankh sign:

`/y25` to set the hieroglyph at 25% in y direction of the cadrat
`/s50` to reduce the size to 50%
`/x18` to set the hieroglyph at 18% in x direction
`/c100` to set colour to grey (three RGB components = 100)

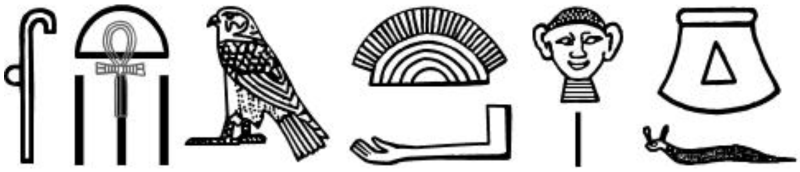


Figure 7. Inscription on the Stela of the king Kamose

(2) Astronomic ceiling (Ramesseum). The full ceiling has been coded and the obtained rendering was given in fig. 6. Let us explain the coding used for one element in the top left register:

[13] The MdC88 syntax is fully supported.

[14] The values refer to the 'base unit', which is a parameter expressed in pixels.

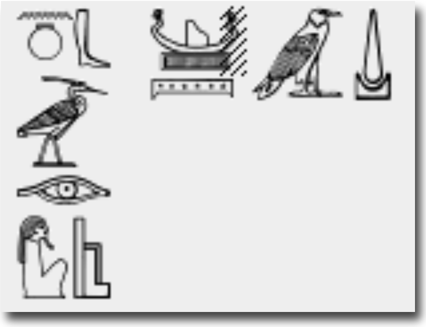


Figure 8. Fragment of astronomic ceiling in Ramesseum

The underlying code is as follows `#poH;d='r' U28 G1 P34/ar25:pt #poV D58*(N35:W24) G31 D4 Q1*A40`. From left to right we see:

`#poH;d='r'`

`oH`: sets orientation to horizontal

`d='r'`: sets direction of writing to right to left

`P34/ar25`: `P34` with `/a` modification operator and `r25` parameter.

`/ar25`: 25% of the hieroglyph is hashed from right side

`#poV`: sets orientation to vertical. Direction coded above is maintained.

3.6. *Conclusion*

The Manuel de Codage, last updated in 1988, needs to evolve but must remain a standard to allow exchanges between Egyptologists to be conducted as easily as possible.

This paper suggests directions for development (new functionalities and updated syntax), proposes principles for these developments, and finally presents how the Rosette Project implements those principles. It now seems relevant to setup a working group commissioned to:

- define new principles of syntax and, as a consequence, list deprecated elements of the MdC88 syntax;
- list ‘basic’ and a first set of ‘extended’ functionalities (see §3.3);
- determine the associated syntax elements;
- determine appropriate milestones for the implementation of above elements.

The MEKETREpository

A Collaborative Web Database for Middle Kingdom Scene Descriptions

Christian MADER, Bernhard HASLHOFER & Niko POPITSCH
Vienna

1. INTRODUCTION

Middle Kingdom (MK) tombs and tomb decorations offer a variety of complex and multi-layered information. However, comprehensive publications that deal with MK scene representations, iconography and scene development are still rare. There is especially a lack of literature performing comparative research on iconography in the MK. In 1922 Luise Klebs published the first assessment of MK representations[1] and in 1978 the last volume of Jacques Vandier's "Manuel d'archéologie égyptienne"[2] appeared. In this publication, he grouped various scenes according to their contents and tried to trace chronological developments in style and iconography. A large quantity of publications dealing with the art of the MK has appeared during the last 40 years, forming an excellent basis for further (comparative) research on scene iconography. However, this tremendous amount of literature poses great challenges to scientists and scholars in the domain of art-history, both in terms of gaining access to publications and keeping up with material being continuously published.

Within the scope of the MEKETRE project, we are developing a specialized software application that will enable scholars to describe MK scenes and scene fragments in a collaborative manner and provide comprehensive search and discovery mechanisms for accessing these items. This application will be referred to as the MEKETREpository, a digital repository of MK art items. It will allow Egyptologists to describe MK items in a structured way and aim at establishing vocabularies for that domain in order to support and improve communication among scholars. The repository and the collected data is already publicly accessible on the Web (http://www.meketre.org) and thus it seeks to make a valuable contribution to future (comparative) research on the MK.

2. METHODOLOGY

Information technology can support Egyptological research in various ways. In the MEKETREpository our focus is on the following four use cases (requirements):

[1] *Die Reliefs und Malereien des mittleren Reiches. (VII.-XVII. Dynastie ca. 2475-1580 v. Chr.) Material zur ägyptischen Kulturgeschichte*, Heidelberg, 1922.

[2] *Manuel d'archéologie égyptienne*. Tome IV. *Bas-reliefs et peinture. Scènes de la vie quotidienne. I^{re} partie: Les tombes*, Paris, 1964; Tome V. *Bas-reliefs et peinture. Scènes de la vie quotidienne. 2ème partie: Élevage, chasse, pêche, navigation*, Paris, 1969; Tome VI. *Bas-reliefs et peinture. Scènes de la vie agricole à l'Ancien et au Moyen Empire*, Paris, 1978.

(1) Storage and retrieval of descriptive metadata for each item.
(2) Collaborative annotation of art items to stimulate cooperation between departments and individual researchers.
(3) Tools for developing and maintaining domain-specific vocabularies.
(4) Assignment of detailed bibliographical references to art items and their details.

When building software for a very specialized group of users, it is not only the implementation of these requirements that will at the end attract the user to the system. With the rise of the so-called "Web 2.0"[3] and wide patronage of websites like Youtube and Facebook, user-supplied content and social networking have attracted users who previously did not have any interest in using the Web. This gives us the opportunity and motivation to build an application that allows users to collaboratively collect and describe fragments of MK tomb decorations. Providing a Web application has advantages for inexperienced users because no local installation is required and the system is accessible from everywhere (when connected to the Internet) and from every device capable of running a Web browser (e.g. also from mobile phones). Since many users are already accustomed to using common Web applications, the technical competence for using the MEKETREpository application is expected to be quite low.

2.1. *Data Model*

The MEKETREpository enables the detailed description of two-dimensional *art items* (cf. Figure 1). Users can create new art items in the repository and may specify:

- the category this item belongs to (see §2.1.1);
- the tomb it belongs to;
- the current location (in situ or some other site, e.g., a museum);
- the position of the item in the tomb (plain language form possible);
- the execution style of the item (e.g., relief, painting, drawing).

Additionally, the user can specify detailed information about *tombs*:

- the necropolis where the tomb is located;
- the tomb number;
- the date;
- the tomb owner.

For both art items and tombs, it is possible to additionally specify the following information:

- a description in plain language;
- keywords (see §2.1.2);
- images depicting the item;
- annotations (see §2.1.3).

By providing this information, it is possible to connect art items to the tombs they originally belong to and describe them in plain language and terms taken from controlled vocabularies.

[3] A definition and comprehensive explanation of the term can be found in O'Reilly, T. 2007. What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software, in: *International Journal of Digital Economics* 65, p. 17-37.

2.1.1. Categorization

There exist many categorization schemes in the Egyptological domain, designed and used by different researchers. In the scope of the MEKETRE project, a new categorization scheme for MK art items will be developed. While this categorization scheme is envisioned to have an independent existence, it will also be linked and cross referenced to Egyptological schemes used elsewhere.

The MEKETRE categorization scheme is a taxonomy developed in a bottom-up, collaborative fashion by the researchers working on the MEKETRE project. Bottom-up means that whenever a researcher enters a new item and requires a new category that is not available in the current scheme, she is able to add this category easily. Depending on the scenes and activities depicted, items can be assigned to one or more categories. The categorization scheme is capable of describing the items contained in the repository but is no generic scheme for MK scenes. Such a generic scheme may be developed in a subsequent process.

2.1.2. Thesaurus

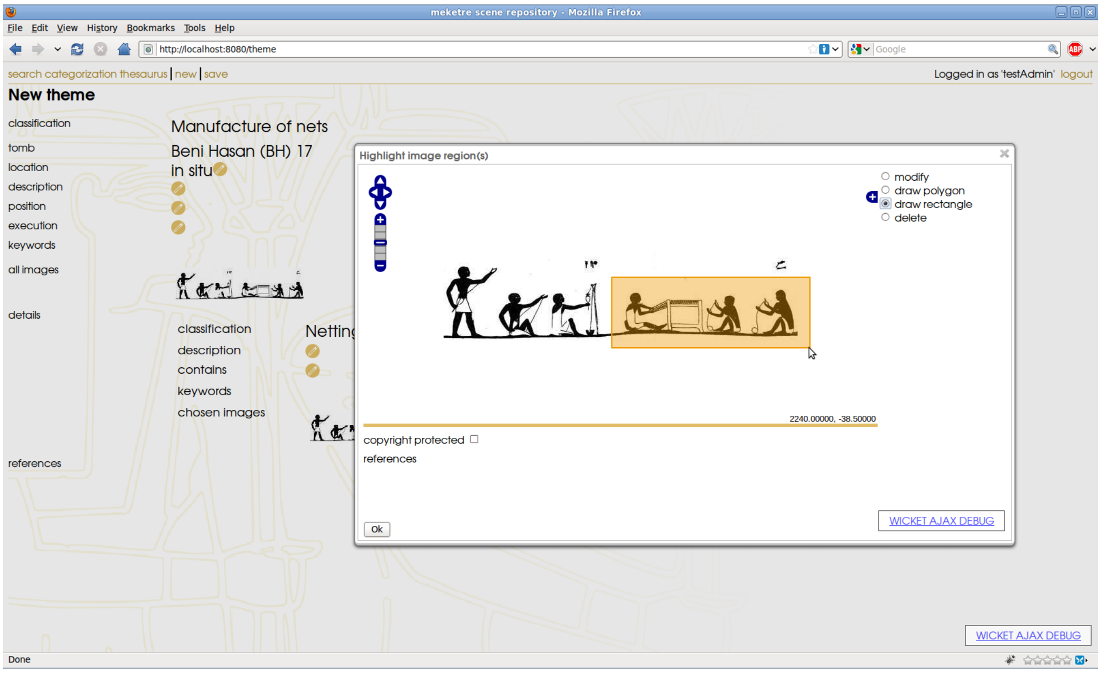
When describing an item, it is beneficial to use terms from controlled vocabularies in order to manage the available data more efficiently. By means of a thesaurus it will be possible to consider the semantic relationships between terms in search and retrieval. Search results could, for example, automatically include items that are tagged with more general terms than the one actually searched for and thus adjust the ranking of the results to produce more useful output. It is important to note that in a single research domain there may exist numerous thesauri, each developed by a different group of researchers, which in their most fundamental form are used to consistently describe items of research within a project. Researchers outside the group often don't have access to such a thesaurus and therefore are unclear about the meaning of its terms. Therefore it is essential to use existing standardized thesauri wherever possible to describe items. Hence, we use the "Multilingual Egyptological Thesaurus" (MET)[4] as a basis for the MEKETRE thesaurus. In cases where the MET does not provide appropriate terms, we allow users to add their terms and the relationships between those terms to a MEKETRE-specific thesaurus. By making this thesaurus publicly available on the Web and linking its entries to the MET it is expected to be a valuable contribution to the Egyptology domain and a complement to the MET.

2.1.3. The Concept of Annotations

In the user interface of the MEKETREpository it is possible to add so-called "details" to an art item or a tomb. A detail is a metadata description for one special aspect of an item. In the computer science community, the term "annotation" has been coined for that kind of architectural pattern. In the MEKETREpository application, each item can have an unrestricted number of annotations and annotations can also refer to each other. Thus it is possible to define relationships between different items like "this scene is contained in the picture of this tomb's wall" or classify parts of an image on a more detailed level.

We carefully designed the user interface of the MEKETREpository to facilitate the annotation of items and to allow for the quick addition of as many annotations to an item as the user desires. The screenshot depicted in Figure 1 illustrates the annotation of a scene item.

[4] http://www.ccer.nl/apps/thesaurus/index.html.


meketre scene repository - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/theme
Google
search categorization thesaurus | new | save
Logged in as 'testAdmin' logout
Highlight image region(s)
Ok
WICKET AJAX DEBUG
WICKET AJAX DEBUG
Done

New theme

classification: Manufacture of nets
tomb: Beni Hasan (BH) 17
location: in situ
description
position
execution
keywords
all images
details
classification: Netting
description
contains
keywords
chosen images
references

○ modify
○ draw polygon
● draw rectangle
○ delete

2240.00000, -38.50000

copyright protected □

references

Figure 1. Adding an annotation to a scene item in the MEKETREpository[5]

2.2. *MEKETREpository Items as Linked Data on the Web*

The World Wide Web (most commonly just referred to as "the Web") was originally designed to support browsing through a large amount of interlinked documents. Although modern Web applications hide this fact to a great deal, working with the Web is essentially a sequence of sending a request for a document (i.e., entering an address or clicking a link) and getting an appropriate response (the actual document or some error message). The documents received by browsing the "traditional" Web are perfectly readable to human users but they are virtually useless for automated processing by computer systems. Efficient data processing requires structured typed data and the HTML markup language most Web pages are written in, does not fulfill this requirement.

Recently, Linked Data[6] has evolved as a method of exposing and linking structured data on the Web. It forms the foundation of the "Web Of Data". Just as the traditional Web serves HTML documents for human users, Linked Data is a method for serving machine-readable RDF data. A Linked Data entity holds data in a machine-readable format (RDF[7]). Furthermore, as the name implies, Linked Data resources are interlinked with resources from other sources, just as a usual Web page contains links to other pages[8]. Other than simple links between regular Web pages, links between RDF items (resources) are typed, i.e., each link has a given semantics and further describes the resource.

A Linked Data resource can describe virtually anything. It can be for example a future event, a concept, or even a feeling. More trivially it may describe an item of a specific domain, for example a person and its properties (e.g., name, age, birthday), a book (e.g., title, author, year of publication) or, in the case of the MEKETREpository, an art item, categorization scheme or thesaurus. These resources are published on the Web, identified by their unique address (URI – Universal Resource Identifier), and can easily be accessed and linked to other resources in the same or in other Linked Data sets. This

[5] In this example, the user has selected the three people on the right of the scene in order to add a description of this particular detail. An unrestricted number of annotations of a scene are supported by our application.

[6] http://www.w3.org/DesignIssues/LinkedData.html.

[7] Resource Description Framework, a W3C standard model for publishing data on the Web. A good source for further reading is http://www.w3.org/RDF/, the complete specification can be found at http://www.w3.org/standards/techs/rdf#w3c_all.

[8] An introduction can be found at http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial.

makes it possible, for instance, to link the Linked Data resource that represents an art item stored in the MEKETREpository with another online available resource (such as a Wikipedia article describing the discoverer of this art item) that is also exposed as a Linked Data resource (as is the case for Wikipedia articles in the DBpedia datasets).

In this fashion, data sets that are available as Linked Data become interlinked over time, forming a huge network of linked resources that can be exploited to learn about related information. Our application could, for example, follow a link to DBpedia, automatically retrieve biographical information about a particular person and display this data next to the depictions of the relevant art item.

2.2.1. Linked Data and the MEKETREpository

We intend to adopt the principles of Linked Data in our MEKETREpository and plan to (i) reuse data from existing Linked Data sources and (ii) publish the data available in MEKETRE as Linked Data on the Web. This will in turn allow other applications to reuse the data collected in MEKETRE by simply addressing the (interlinked) MEKETRE items by their URIs. More specifically we want to publish structured controlled vocabularies using the Simple Knowledge Organization System (SKOS[9]). For describing item metadata we will investigate the applicability of existing metadata standards such as Dublin Core (DC[10]), Friend-of-a-Friend (FOAF[11]), and others. One of the goals of the MEKETREpository application is to contribute to the Linked Data cloud[12] and provide interoperability with other data sources following the same standards.

2.3. *Copyright Issues*

Attaching media objects such as images to collected art items or tombs is useful for users to obtain a quick overview of the material or of interesting items of a search result. However, some pictures of MK scenes or scene fragments that should be added to the MEKETREpository's database are only available in books that are still protected by copyright. Hence the MEKETREpository allows the creation of items without any attached image. Such items are basically metadata descriptions that characterize a real-world art item but lack its depiction.

As an alternative, it is possible to tag an image as copyrighted when uploaded. This prevents image access for users who are not logged into the MEKETREpository application or do not have the right to view copyrighted material. As a general rule-of-thumb we propose to upload public domain material when available. Since any number of images can be uploaded, it is also possible to upload copyrighted material together with non-copyrighted material illustrating the same item.

2.4. *Accessing the Repository*

We designed the MEKETREpository for two different types of users. The first type relates to human users like researchers and students. Researchers will have read- and write-access to the repository and provide material along with a qualified description of the content. Students may browse the collected data for comparative research purposes without contributing to the repository. For both researchers and students we provide an easy-to-use Web application for accessing the stored items and performing their work.

Machine users represent the other type of user for the MEKETREpository: other systems connected to our repository via the Web with read-access to the stored data. Since we provide Linked Open Data (as described in §2.2), which is a novel approach for publishing machine-readable data,

[9] http://www.w3.org/2004/02/skos/.

[10] http://dublincore.org/.

[11] http://www.foaf-project.org/.

[12] To get an impression, see http://richard.cyganiak.de/2007/10/lod/.

our data can easily be queried and integrated with other data sets, potentially originating from domains outside of Egyptology.

2.5. *Long-term Archiving*

Preserving the collected data using a Long-Term Archiving Solution (LTAS) is another aim of the project. We are not going to develop a new solution, but focus on how to integrate the MEKETREpository with an already existing LTAS.

The University of Vienna hosts a digital asset management system with long-term archiving functions called PHAIDRA.[13] It is based on the popular Fedora Commons Repository Software[14] and can hold any kind of digital object, available worldwide around the clock with continual citability. PHAIDRA also uses metadata to store the content but its metadata standards are fixed and not easily tailorable to domain-specific needs. In the context of the MEKETREpository project, we use PHAIDRA as an additional storage solution. The data stored in the MEKETREpository is periodically replicated to PHAIDRA.

PHAIDRA is a solution for general archiving purposes whereas the MEKETREpository is custom-tailored to be used by Egyptologists. Since MEKETRE is an interdisciplinary project we are working hand-in-hand with our colleagues from the Institute of Egyptology to provide them with the tool they need to perform their research. The workflow in the MEKETREpository is optimized for finding, browsing and comparing scenes. Our strategy is to combine the intuitive user interface of the MEKETREpository with the long-term data archiving capabilities of PHAIDRA. Since the MEKETREpository also works with common metadata standards, the conversion to PHAIDRA datasets should be straightforward for most entries. The two systems will exist side-by-side and can be queried independently. It is important to note that the MEKETREpository is designed to exist on its own but we decided to additionally replicate the data to PHAIDRA to make use of an existing well-proven long-term archiving repository with relatively low effort expenditure.

2.6. *System Architecture*

Since the MEKETREpository application is expected to be used and maintained beyond the project's three-year limit, it is important to implement it using industry-standard components that are available under an open source license. The system is designed as a three-tier application consisting of a persistence layer, an intermediate service layer and a user interface layer. The main programming language is Java and the user interface layer is implemented using the Apache Wicket Web framework[15]. Operations on the data are passed to the service layer which in turn utilizes the persistence layer to save data to or obtain data from a relational database. The service layer is also responsible for providing Web services for other systems querying data. The persistence layer uses the Java Persistence API (JPA) to describe the data model. Hibernate is used as the JPA provider together with a MySQL database storing the actual data. Uploaded images of items and tombs are stored in the file system and are managed by an IIP (Internet Imaging Protocol[16]) Image Server which is a FCGI Application running on an Apache webserver. For fulltext search we make use of the Solr search server that runs as a separate process. Managing literature is done using an existing Web application (refbase[17]). However, by utilizing refbase's OpenSearch[18] interface, it is possible to edit literature

[13] https://phaidra.univie.ac.at/.

[14] http://fedora-commons.org/.

[15] http://wicket.apache.org/.

[16] The specification can be downloaded from http://iipimage.sourceforge.net/IIPv105.pdf.

[17] http://www.refbase.net/.

[18] http://www.opensearch.org.

references directly from the MEKETREpository's user interface. An overview of the system and the involved components is shown in Figure 2.

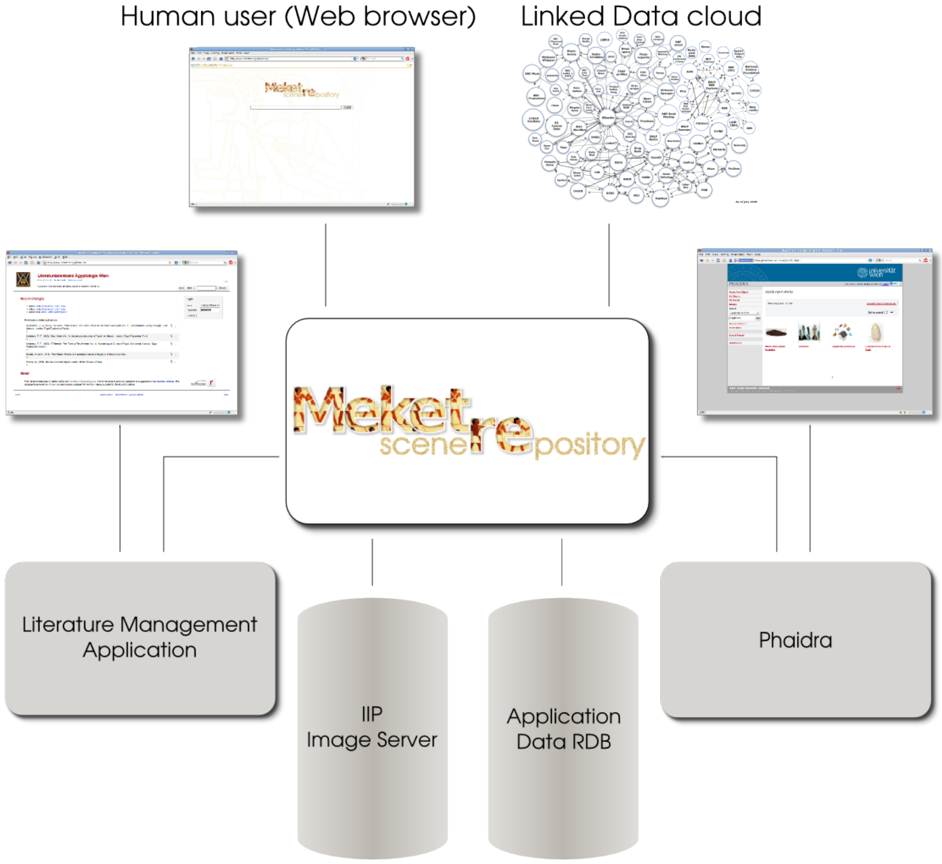


Figure 2. The MEKETREpository and its components[19]

3. Related Work

The Oxford Expedition to Egypt (OEE)[20], which is affiliated academically to Linacre College, University of Oxford, created a scene details database. In the years from 2003 until 2006 the expedition collected data of scene details in Old Kingdom (OK) monuments. The database[21] went online in 2007 and is now freely available. It has been developed in close collaboration with the Archaeology Data Service (ADS)[22]. Similar to the categorization scheme developed during the work on the MEKETREpository's content, the OEE database uses a hierarchical scheme to organize the data into "Themes", "Scene types" and "Scene details". We plan to adopt and integrate already existing data modeling practices developed at ADS and give feedback in order to ultimately build a basis for publishing scene descriptions as Linked Data on the Web.

Scenes and their accompanying texts in OK tombs are also covered in the scope of the Leiden Mastaba Project (LMP)[23], also known as "MastaBase". In contrast to the MEKETREpository the data is not published directly on the Web for public access but purchasable on CD-ROM. There are also restrictions on the types of computer systems that may use this database, further restricting the possible user base of these data.

[19] Both the Literature Management Application and Phaidra can be queried independently from the MEKETREpository using a Web browser.

[20] http://www.oxfordexpeditiontoegypt.com/index.html.

[21] http://ads.ahds.ac.uk/catalogue/archive/oee_ahrc_2006/.

[22] http://ads.ahds.ac.uk/.

[23] For a project summary see http://www.peeters-leuven.be/boekoverz.asp?nr=8170.

4. CONCLUSIONS AND FUTURE WORK

The MEKETREpository is a software solution capable of describing MK two-dimensional art items at an unrestricted level of detail. The contained data is published as Linked Data on the Web utilizing controlled vocabularies.

We aim at making it as easy as possible for scholars to enhance existing and develop new vocabularies. Since these vocabularies are published as Linked Data on the Web, it is essential to provide mappings to other existing vocabularies (e.g., DBpedia), so a strategy must be developed to create these mappings.

Since we expect the MEKETREpository to be used by more than one researcher concurrently, we will investigate possibilities for Web-based collaborative thesaurus editing. This brings up a whole bag of new challenges, namely how to track and record changes. For example, when one user deletes or reorganizes a categorization while another user is just about to use the same term/concept in a new annotation, a conflict occurs that has to be resolved. Furthermore, the changes in the vocabularies need to be tracked for documentation reasons and to provide the basis for further discussion. Annotating items is done collaboratively and the editing of the vocabularies is likewise a collaborative process.

Another challenge is the development of a user-friendly interface that allows users to formulate sophisticated queries on the data. The MEKETREpository forms the basis of further research, so it is essential that users are able to navigate through the data efficiently. Our users are not expected to have a strong background in computer science, thus we cannot expect them to use a specialized query language (e.g., SPARQL) to search for items of interest. Approaches for accessing data in a way that conforms to the Web 2.0 user interface paradigm will have to be investigated and integrated into the application.

Based on the relationships between items and annotations it would be possible to generate new visually appealing views of repository items and their links to one another. These views should help both scholars and researchers to get an overview and track interrelations between various items. We will further research the possibility of automatizing the creation of these methods of information visualisation.

The Digital Puzzle of the *Talatat* from Karnak

A Tool for the Three-Dimensional Reconstruction of Theban Buildings from the Reign of Amenhotep IV

Nathalie Prévôt

Institut Ausonius (CNRS-Université de Bordeaux)

1. The theban *talatat*

The *talatat* from Karnak, or more precisely the usage of computers for the reconstruction of Atonist temples built with these *talatat* during the reign of Amenhotep IV-Akhenaton, constitutes a recurring theme in the context of the colloquia "Informatique et Égyptologie".[1] In the first session in 1984, Robert Vergnieux[2] gave a lecture on the database called *talatat*, and on his intentions and ambitions to put it online. It is now available on the Internet with protected access, and will be open to the public at the end of 2011.

The subject of *talatat* is far from being exhausted, and many aspects of its study still remain to be discussed: the conditions of their reconstruction, the restoration of their cohesion, as well as their relevance, significance, etc. Before presenting the tool of the interactive digital puzzle, which gives this article its title, it is necessary to contextualize it and so explain quickly the ATON-3D program which made its development necessary.

2. The ATON-3D project

We are interested in a very methodological perspective of the *talatat* thanks to the revival of research on Atonist temples from Karnak, namely a vast interdisciplinary and international research project implemented in 2009 with the help of the French national agency for research (ANR-08-BLAN-0202-01). This project, called ATON-3D, aims to study the architectural policy of Akhenaton, both in Karnak and in Amarna, thanks to the tridimensional digital modelling of structures built during this reign. For Amarna, we have a lot of sources: numerous excavation reports and many reliefs from Amarna tombs are very informative, together with the tremendous advantage that Atonist temples have not been hidden by later constructions. On the other hand, most of the Amarna *talatat* disappeared, because they were in limestone and thus where burned in lime kilns during the Middle Ages.

In Karnak, there are tens of thousands of sandstone *talatat* which constitute the most important sources of documentation for the ATON-3D project. All the Atonist constructions were dismantled in the post-Amarna period, and the evidence was unfortunately scattered. We know the names of buildings, attested in some few texts found on preserved *talatat*, but we do not know their functions, plans, or locations.

[1] Brocard 1994; Vergnieux 1985a, 1988, 1990, 1994.

[2] Vergnieux 1985b: 223.

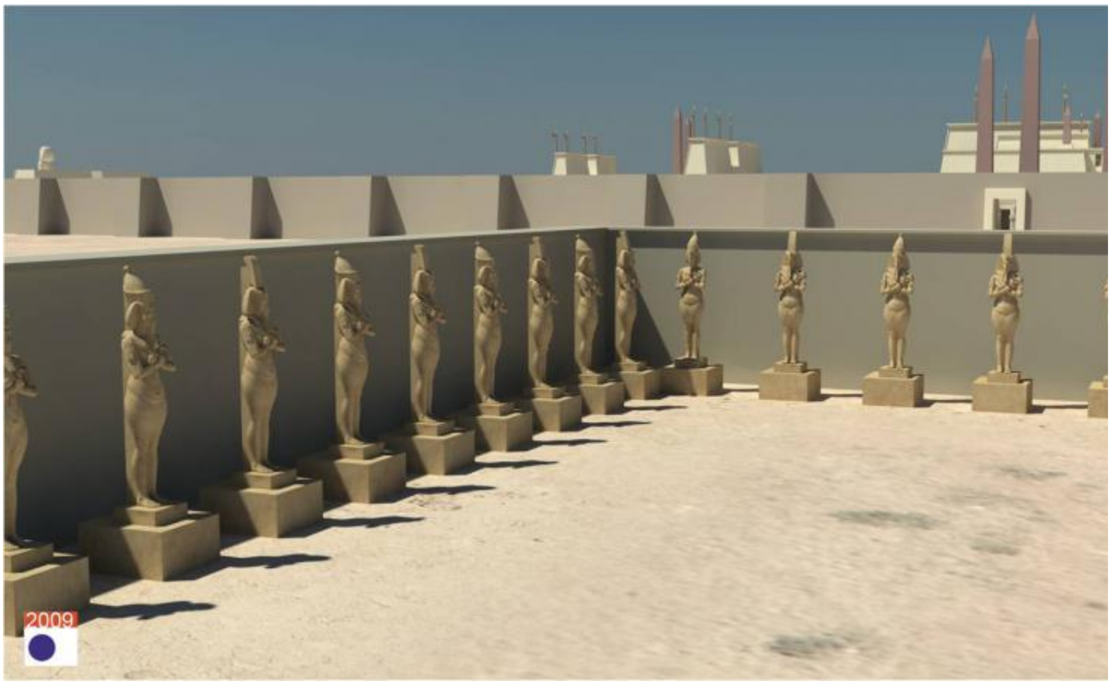


Figure 1. First results of the ATON-3D program:
The gem-pa-aten outside the boundaries of the Karnak precinct of Amon-Re (East; WIP)

Among the approximately 12 000 *talatat* which have been extracted from the western wing of the ninth pylon by the *Centre Franco-Égyptien d'Études des Temples de Karnak* (CFEETK), Robert Vergnieux has already systematically studied 6 666 stones of the last 24 courses in the base of the pylon.[3] However, it is necessary for us to review them today from the perspective of the three-dimensional reconstructions made by the ATON-3D project. Their importance is twofold. First, they constitute the walls of buildings which we want to restore, thus the more of it we assemble, the more we obtain dimensions, widths and heights of walls. Second, as the ornamental scheme of buildings erected with these *talatat* contains decorative scenes representing temples which were functioning during this period, the more of it we assemble, the more we obtain representations of the structures we want to restore.

3. THE *TALATAT* CORPUS

The very first database was created in the 80s on an Apple II. This database referenced 12 000 blocks which were distributed in about forty courses inside the ninth pylon. We were able to obtain the backup made on 5 ¼ inch floppy disks containing the corpus studied in Robert Vergnieux's thesis,[4] namely the lower layers of re-use (from the 24th to the 39th). We have recovered this database,[5] 25 years after the floppies were recorded, to ensure that this data be preserved and made accessible in the long term.

[3] Vergnieux 1999.

[4] Vergnieux 1999.

[5] We had to find an Apple II and floppy disks with the programming language Pascal, because Apple II must be started under the system Pascal on floppy to be able to read the diskettes containing the old *talatat* database. The Apple II having naturally neither USB ports nor network connections had to be connected via a serial cable to a laptop still equipped with a COM1 serial port; the "modem" port of the Apple II (DIN 5 pins) and the COM1 port of the PC (V11 9 pins) were connected by a crossed cable ("null modem") made for that purpose. The application "Hyperterminal" was launched on the PC and configured to record on the hard disk all the traffic from the serial interface COM1. Finally, we used the file manager program in order to send the data to the "modem" port.

Today, with the beginning of the ATON-3D program, the *talatat* database has evolved a lot to allow integration into our information system ArchéoGRID[6] (fig. 2) and has been published online, which is a guarantee of preservation. This corpus is stored in a secure data centre[7] with a strong human and material infrastructure that will now ensure the conservation of images and metadata of the *talatat*.[8]

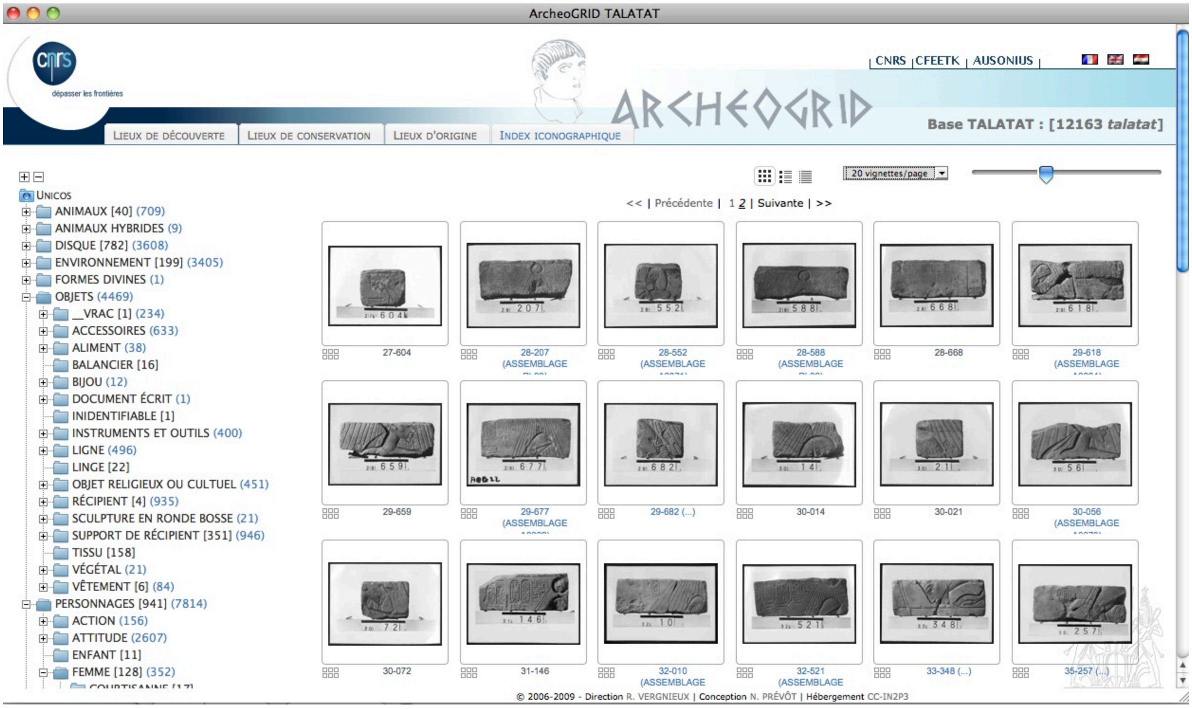


Figure 2. Archeogrid-Talatat

In partnership and in agreement with the CFEETK,[9] the *talatat* database is today available for all those involved in the ATON-3D project and should be publicly accessible in 2012.

The current database contains a little more than 12 000 documents. It is completely extensible, and we hope to add to it all other *talatat* found in Luxor, Ermant, Tod, Medamud, Hermopolis, etc. as well as establish links with other databases[10] also containing *talatat*.

For the documentation studies and the metadata entries, we use a French/English multilingual thesaurus. Each document is defined by descriptors of identification (inventory number, former

[6] Archéogrid is the Information System developed by the Plateforme Technologique 3D in the Institut Ausonius, which allows the linking of the 3D reconstruction of a building or of an archaeological object to the heterogeneous information sources (photos, excavation reports, surveys, architectural drawings, historical iconography, etc.) that allowed its representation.

[7] ADONIS-tge is a Major Facility (Très Grand Équipement) launched by the CNRS (French National Centre for Scientific Research) open to other partners. It aims to promote integrated access to digital data and documents in the Humanities and Social Sciences: "la grille Adonis". It offers grid facilities (computing, storage, Web services, virtual environment, tools) and is organized within the Computing Centre of the National Institute for Nuclear Physics and Particle Physics (CC-IN2P3).

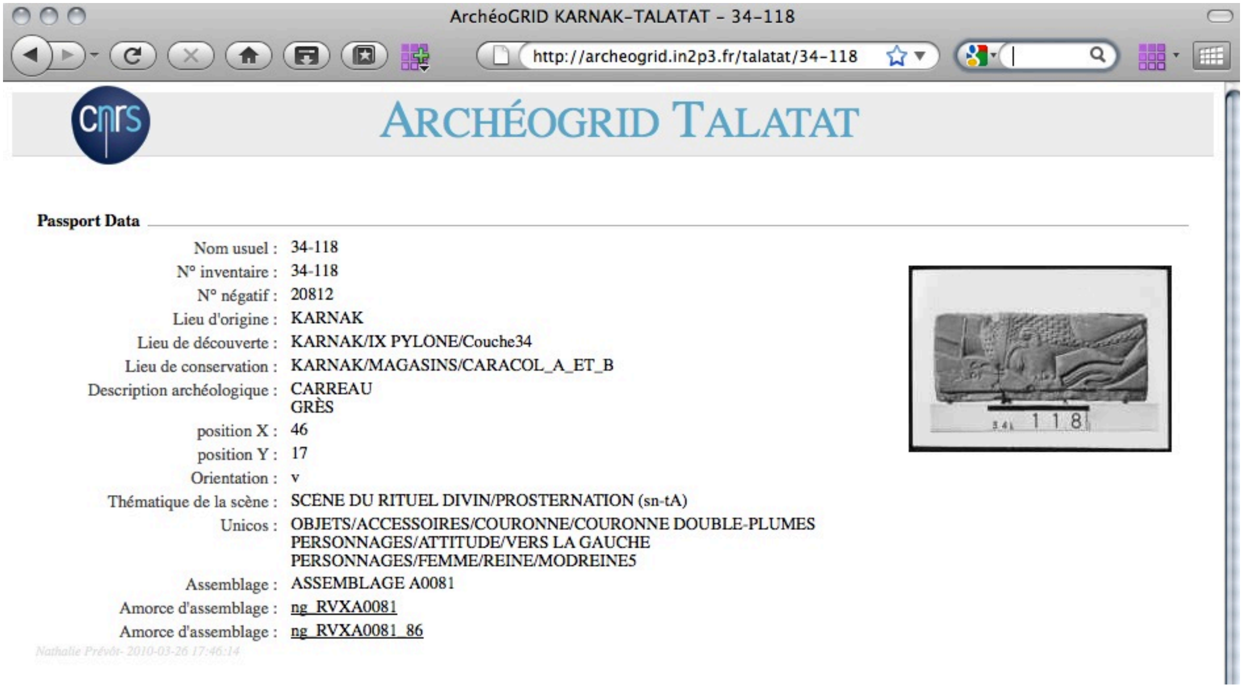
[8] The engineers of the CC-IN2P3/TGE-Adonis perform with daily, weekly, monthly copies of security. Additionally, they proceed with implementation of warning devices of type technology and economic watch and with regular upgrades in order to ensure that we use software that are not only free and open source, but correspond technically to the state of the art. Moreover, they encourage us to respect systematically the standards.

[9] http://www.cfeetk.cnrs.fr/index.php?page=axe-6.

[10] For example, a few months ago, the ARCE presented to us a new database containing 16 000 *talatat* stored in the storehouse leaning against the temple of Khonsu in Karnak. Since 2009, all these blocks have been restored, photographed and documented. It is of the highest interest and importance to interconnect these two databases.

inventory numbers), of location (place of origin, place of discovery, place of conservation), of descripttion (archaeological, iconographic, epigraphic), and is accompanied by bibliographical information.

To avoid the subjective nature of the iconographic indexing, we opted for an analytical description, incorporating the concept proposed by Robert Vergnieux of knowledge representation using *unicos*.[11] The aim here is to separate different types of information which we want to be able to query separately. A *unico* is the iconographic unity which corresponds to an independent iconic visual sign, e.g. human beings (royal family, priests, soldiers, courtiers, foreigners, artists, etc.), animals (cattle, horses, etc.), products of human activity (offering tables, thrones, ships, architecture, etc.), nature (sun-discs, canal and water, trees, flowers, etc.). Describing a *talatat* consists of enumerating the *unicos* present on its decorated face.



ArchéoGRID KARNAK-TALATAT - 34-118

http://archeogrid.in2p3.fr/talatat/34-118

ARCHÉOGRID TALATAT

Passport Data

Nom usuel :	34-118
N° inventaire :	34-118
N° négatif :	20812
Lieu d'origine :	KARNAK
Lieu de découverte :	KARNAK/IX PYLONE/Couche34
Lieu de conservation :	KARNAK/MAGASINS/CARACOL_A_ET_B
Description archéologique :	CARREAU GRÈS
position X :	46
position Y :	17
Orientation :	v
Thématique de la scène :	SCÈNE DU RITUEL DIVIN/PROSTERNATION (sn-tA)
Unicos :	OBJETS/ACCESSOIRES/COURONNE/COURONNE DOUBLE-PLUMES PERSONNAGES/ATTITUDE/VERS LA GAUCHE PERSONNAGES/FEMME/REINE/MODREINE5
Assemblage :	ASSEMBLAGE A0081
Amorce d'assemblage :	ng_RVXA0081
Amorce d'assemblage :	ng_RVXA0081_86

Nathalie Prévôt- 2010-03-26 17:46:14

Figure 3. Metadata of a *talatat*

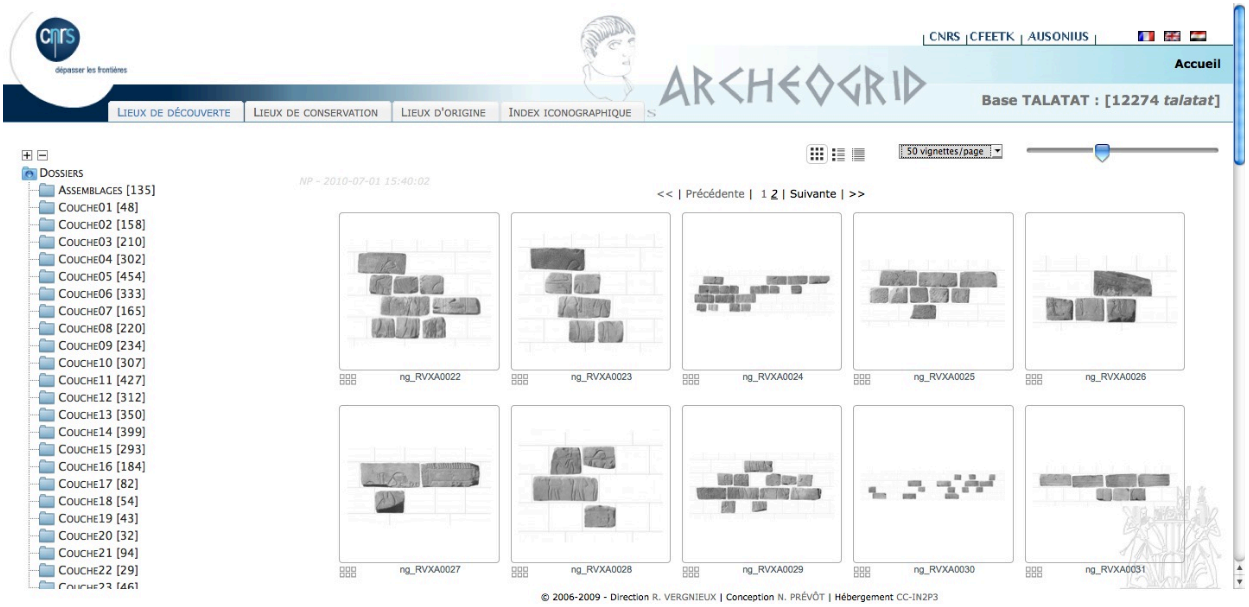
4. THE *TALATAT* PUZZLE

Talatat, if taken separately, are information poor. We have to reassemble them in order to get information that allows us to reconstruct the buildings.

We have developed a tool that facilitates not only the reconstruction process, but also the storage and archival of the data, allowing them to become in turn research documents that enrich the database. Such a tool has the further benefit of eliminating redundancy of effort — many scenes that have been reconstituted by draftsmen or PhD students working in Karnak using scissors and tracing paper have until now not been exploited for reconstruction purposes.

This collaborative assistant to the assemblage of blocks is interfaced with the corpus of *talatat*. It is a kind of puzzle on screen, which allows the correct placement of blocks thanks to a grid that takes into account the usual construction design of alternating rows of stretchers and rows of headers. Courses of headers are always aligned vertically in parallel. Each stretcher overlaps three headers and rows of stretchers are offset between them vertically by the length of half a header. Hence we ensure that a *talatat* matches laterally only with another one showing its decoration on the same surface as itself (e.g. headers next to headers) and above and below with a block decorated on a different side.

[11] Vergnieux 1990.



Base TALATAT : [12274 *talatat*]

50 vignettes/page

NP - 2010-07-01 15:40:02

<< | Précédente | 1 2 | Suivante | >>

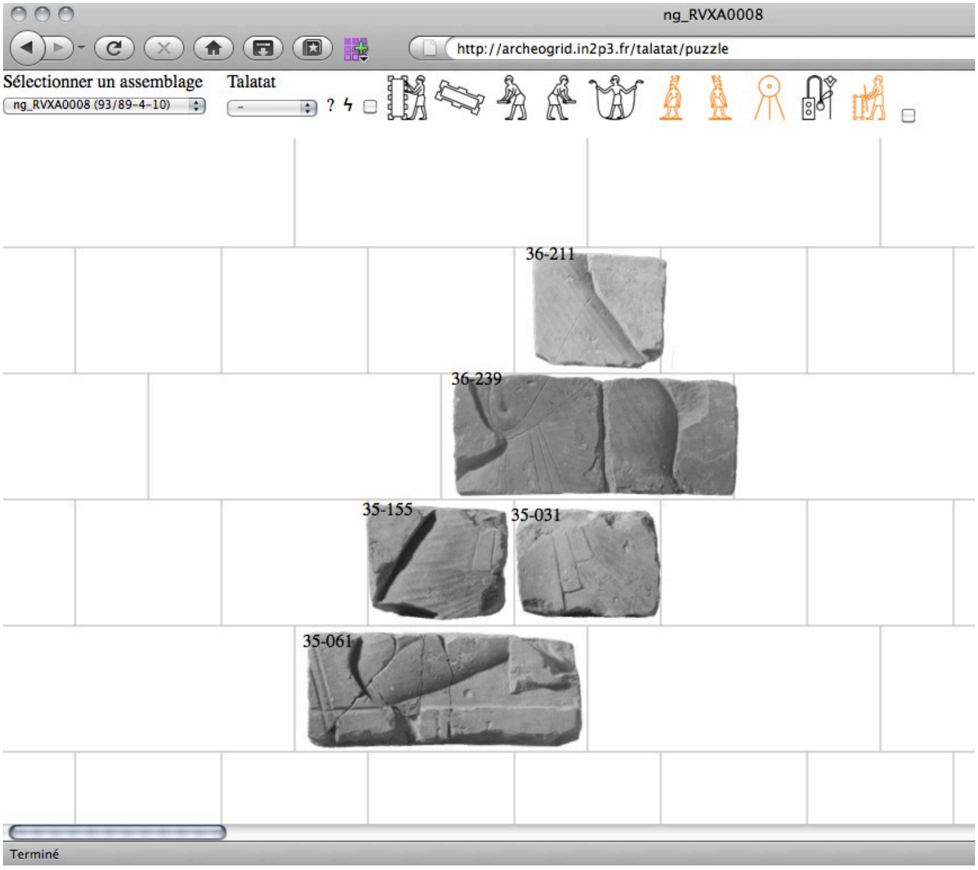
Dossiers
- Assemblages [135]
- Couche01 [48]
- Couche02 [158]
- Couche03 [210]
- Couche04 [302]
- Couche05 [454]
- Couche06 [333]
- Couche07 [165]
- Couche08 [220]
- Couche09 [234]
- Couche10 [307]
- Couche11 [427]
- Couche12 [312]
- Couche13 [350]
- Couche14 [399]
- Couche15 [293]
- Couche16 [184]
- Couche17 [82]
- Couche18 [54]
- Couche19 [43]
- Couche20 [32]
- Couche21 [94]
- Couche22 [29]
- Couche23 [46]

ng_RVXA0022 ng_RVXA0023 ng_RVXA0024 ng_RVXA0025 ng_RVXA0026
ng_RVXA0027 ng_RVXA0028 ng_RVXA0029 ng_RVXA0030 ng_RVXA0031

© 2006-2009 - Direction R. VERGNIEUX | Conception N. PRÉVÔT | Hébergement CC-IN2P3

Figure 4. The assemblages in Archeogrid

On the left part of the screen, we search for potential neighbouring blocks using targeted queries (stretcher or header + the most likely *unicos*) and the results are displayed immediately; on the right part we test whether the block which seems to correspond matches with the others already in place. The photographs of the *talatat* displayed in the puzzle are cropped so that they can be set edge to edge. Of course, the images are represented at exactly the same scale.



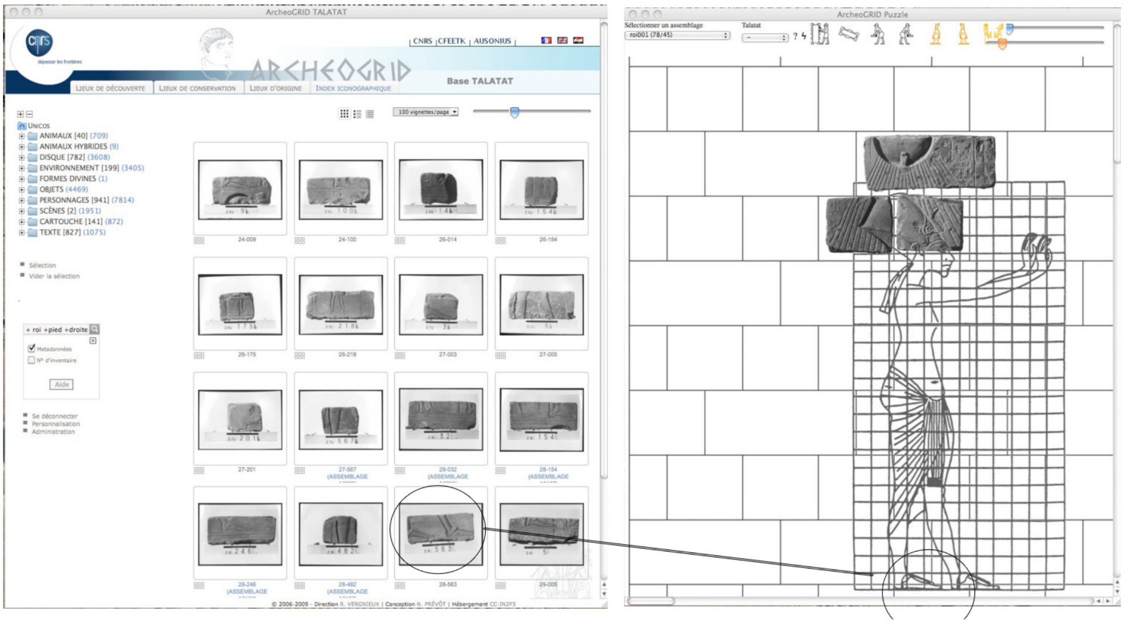
http://archeogrid.in2p3.fr/talatat/puzzle

Sélectionner un assemblage: ng_RVXA0008 (93/89-4-10)

Talatat: -

Figure 5. Assemblage A0008 with display of inventory numbers

Numerous options are available, such as zooming, rotating blocks, multiple selecting to move several *talatat* together, as well as the possibility of applying a layer of gridlines to follow the canon of proportions in the Atonist era (fig. 6). The images thus generated will be used as textures in the 3D models.


Sélectionner un assemblage: roi001 (78/45)

Talatat: -

100 vignettes/page

+ roi +pied +droite

☑ Metadonnées

□ Nº d'inventaire

© 2006-2009 - Direction R. VERGNIEUX | Conception N. PRÉVÔT | Hébergement CC-IN2P3

Figure 6. Using the puzzle tool

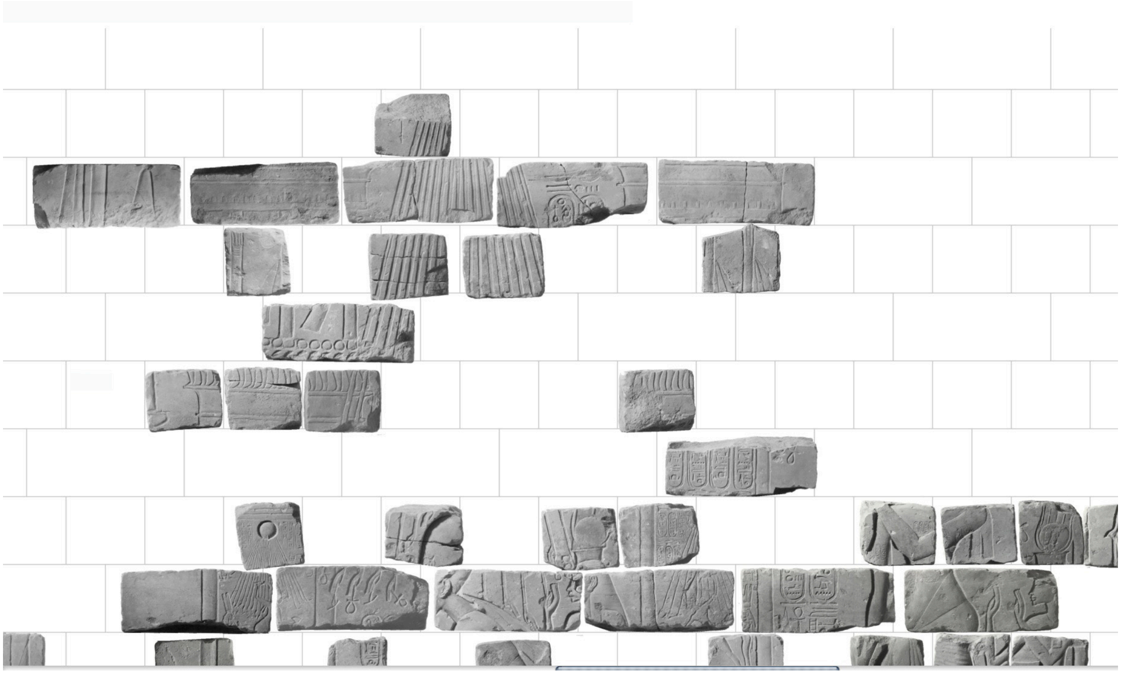


Figure 7. A portion of the assemblage A0011 published by R. Vergnieux, completed with the puzzle tool and with the help of D. Laboury[12]

The maximum height of the wall that it was possible to reconstruct with this tool is 15 cubits, or approximately 8 meters high, which corresponds to 35 courses of *talatat*. The maximum length is about 30 meters.

[12] The reconstruction, still very incomplete, has 113 *talatat*; it is 11 meters long and 3,75 meters high — 17 courses in total.

The system backs up the temporarily reconstituted scenes by storing in the database the positions of every block on the grid. It is thus possible to resume a reconstruction several days after it was drafted. The current work can be even pursued by another researcher tempted by the addition of a likely block. So a number of users can participate in the collective construction of an interpretation.

This development was made possible thanks to the arrival of HTML5[13] and CSS3[14] formats, which allow the development of very ergonomic web applications and the performance of drag & drop operations and rotations of images, etc. This tool is certainly less advanced than Adobe Photoshop, but it does not require users either to download the photos of *talatat* on one's own machine, nor to install any software, providing users have a simple web browser installed.

The association between the puzzle and the database allows the immediate integration of new information in the system based on statistics obtained from the analysis of the number of reassembled stones, the number of missing stones, heights, lengths, number of courses of the reconstituted walls, etc.

Thus the database of *talatat* gets bigger and bigger according to the successful use of the puzzle combined with the subsequently obtained results. Archeogrid displays these data resulting from the research (the results of the reconstructions) together with the sources themselves (= the *talatat*).

5. INTEROPERABILITY OF THE *TALATAT* DATABASE: TOWARDS THE SEMANTIC WEB

In accord with the CFEETK, our corpus has been made machine-readable, with the double goal of making it accessible to researchers of the proto-Amarna period, and sharing it with other teams developing databases of *talatat* found in Karnak or elsewhere. This allows us to distribute resources from different research centers without gathering them physically or needing to duplicate them, which would quickly present difficulties for updating and archiving. In addition, it would be difficult to exploit resources effectively if they were stored in various locations, besides the complications associated with using a number of heterogeneous tools.

We chose to publish the data concerning the *talatat* presented on the Archeogrid website in an existing format so that they are easily reusable and immediately interoperable with data published with the same protocols. Therefore we use RDFa,[15] which allows the insertion of descriptions corresponding to the data model RDF[16] in the HTML representation of a resource. Consulting with a simple web browser the source code of the record of a *talatat* reveals a structured representation of the information according to the principles of RDF with the use of several documentary vocabularies:

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN"
"http://www.w3.org/MarkUp/DTD/xhtml-rdfa-1.dtd">
<html xml:lang="fr" version="XHTML+RDFa 1.0"
xmlns="http://www.w3.org/1999/xhtml"
xmlns:foaf="http://xmlns.com/foaf/0.1/"
xmlns:dc="http://purl.org/dc/elements/1.1/"
```

[13] This is the 5th major revision of the core language of the World Wide Web: the Hypertext Markup Language (HTML) for structuring and presenting content. In this version, new features are introduced to help Web application authors.

[14] Cascading Style Sheets (CSS) Level 3 is a language for describing the rendering of a document written in HTML or XML.

[15] RDFa, which means "RDF in attributes", is a meeting between a document format that represents the resources and a data model used by machines that describes these resources. It thus allows users to make queries on a text document with the standardized query language SPARQL just as they are done in a database with SQL.

[16] As its name suggests, the RDF (Resource Description Framework) allows the description of resources (while the HTML allows the construction of representations of these resources); see Mader *et al.* (current volume) for another application of this technology within the field of Egyptology. Developed by the W3C within the framework of the activities of semantic Web, RDF is not, strictly speaking, a metadata schema. It is a model of description of the structured data inspired by graph theory. Its genericity and its flexibility offer a framework of interoperability for describing all types of resources in a networked environment like the Web.

```
xmlns:cc="http://creativecommons.org/ns#"
xmlns:dcterms="http://purl.org/dc/terms/"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#
xmlns:tal="http://archeogrid.in2p3.fr/talatat/ns/1.0/"
xmlns:xsd="http://www.w3.org/2001/XMLSchema#">
```

The advantage of RDF is that it is possible to exploit the data, whatever the vocabulary used, without having to convert it, unlike XML for which it is necessary to convert the data if the user is not using the same schema. Thus it does not place the imposition on various teams to agree on a single method of structuring metadata, nor does it limit them to a lesser common denominator to ensure interoperability.[17] RDF provides a framework for great flexibility to mix and associate terms from existing vocabularies, as well as invent our own in the combination which is best suited for our particular contents. The interoperability offered means on one hand that our data are exploitable by existing tools, and on the other hand that they can really be connected with other data via the web.

To indicate keywords, we insert "dc:subject" tags that refer to Dublin Core[18] vocabulary and that we link to our thesaurus:

```
<span rel="dc:subject">
<span property="tal:unico">
<a href="[lien vers les unicos]">
   OBJETS/ACCESSOIRES/COURONNE/COURONNE DOUBLE-PLUMES</a>
</span>
```

For the geo-localization of the *talatat*, we use Dublin Core terms with the tag "spatial", which allows us to specify values of latitude and longitude while referencing in the WGS vocabulary validated by the W3C (geo:lat and geo:long).

```
<span rel="dcterms:spatial">
<span property="geo:lat" content="25.716269"></span>
<span property="geo:long" content="32.655174"></span>
</span>
```

The thesaurus used has been entirely converted to the language SKOS,[19] which, like an ontology, aims at an increased interoperability for exchanges of lexicon, the integration of a semantic Web as well as automatic machine processing.

The normalized syntax we use can be exploited in various ways to obtain information from our resources. To do this, it is simply necessary to retrieve the sitemap.xml file[20] at the site root (fig. 8) and to extract the data correctly marked with RDFa using appropriate parsers, whether written in XSLT,

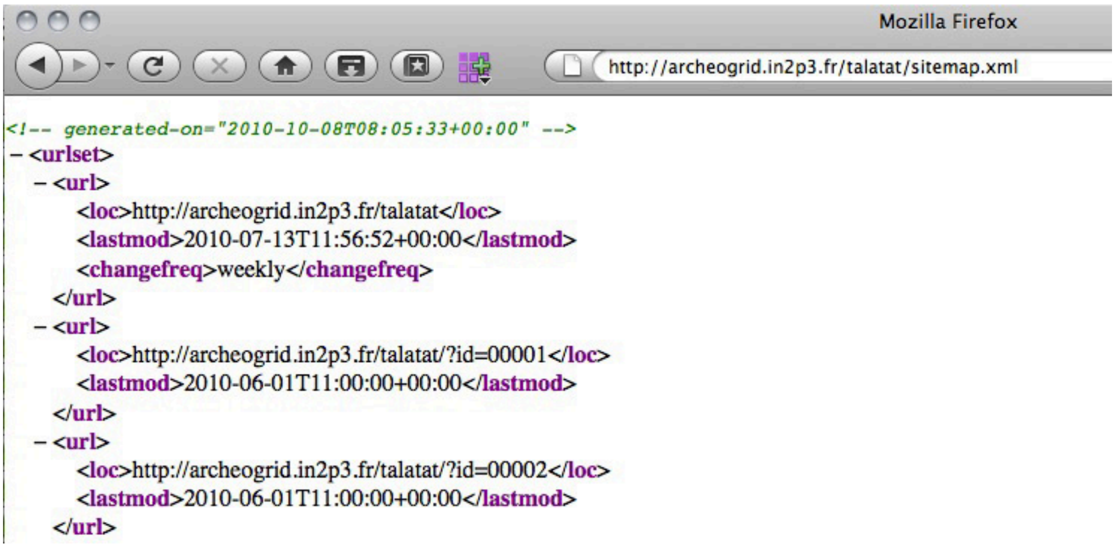
[17] As it is the case with OAI-PMH.

[18] The Dublin Core Metadata Element Set is a vocabulary of fifteen properties for use in resource description. Its elements are broad and generic, usable for describing a wide range of resources.

[19] Simple Knowledge Organization System – The language SKOS designed for representation of thesauri, classification schemes and taxonomies is an ontology allowing on one hand the representation in a multilingual context of every type of structured and controlled vocabulary, and on the other hand the alignment of various vocabularies, the objective being the machine exploitation of resources published on Web.

[20] Introduced by Google, the Sitemaps protocol allows users to indicate to search engines the resources of a website to be indexed. It is an XML file which contains (for every resource): its URL, its date of last modification, the frequency of revision and the relative importance with regard to the other URLs of the site. The use of the Sitemaps protocol allows us to guide the collection of the data and metadata on *talatat* and assemblages.

JavaScript or even Python. With these tools, which already exist,[21] RDF tuples can be recreated, and exploited like any other data in this format, and can then be shown on any other website.



```
<!-- generated-on="2010-10-08T08:05:33+00:00" -->
- <urlset>
  - <url>
      <loc>http://archeogrid.in2p3.fr/talatat</loc>
      <lastmod>2010-07-13T11:56:52+00:00</lastmod>
      <changefreq>weekly</changefreq>
    </url>
  - <url>
      <loc>http://archeogrid.in2p3.fr/talatat/?id=00001</loc>
      <lastmod>2010-06-01T11:00:00+00:00</lastmod>
    </url>
  - <url>
      <loc>http://archeogrid.in2p3.fr/talatat/?id=00002</loc>
      <lastmod>2010-06-01T11:00:00+00:00</lastmod>
    </url>
```

Figure 8. Sitemap.xml

In summary, all the data on *talatat* as well as the research data derived from them, such as the walls rebuilt using the puzzle, are usable and quickly interoperable, without being duplicated or moved, and always in their latest version, while they are continuously documented and interpreted by successive refinements.

Bibliography

Adida, Ben, Mark Birbeck, Shane McCarron & Steven Pemberton. 2008. RDFa in XHTML: Syntax and Processing – W3C Recommendation 14 October 2008, http://www.w3.org/TR/rdfa-syntax.

Billet, Sophie. 1988. Intelligence artificielle et base de données pour lassemblage des talatat, in: *Fifth International Congress of Egyptology, October 29 — Novembrer 3, Cairo, 1988, Abstracts of Papers*, International Association of Egyptologists, Cairo, p. 23-24.

Billet, Sophie, Michel Gondran & Robert Vergnieux. 1988. Intelligence artificielle et 'talatat', in: *Fifth International Congress of Egyptology, October 29 — Novembrer 3, Cairo, 1988, Abstracts of Papers*, International Association of Egyptologists, Cairo, p. 25-26.

Brocart, Yannick. 1996. Quelques exemples de restitutions virtuelles, in: *Informatique et Égyptologie* 10, p. 31-34.

Brickley, Dan & R.V. Guha. 2004. RDF Vocabulary Description Language 1.0: RDF Schema - W3C Recommendation 10 February 2004, http://www.w3.org/TR/rdf-schema.

Miles, Alistair & Sean Bechhofer. 2009. SKOS Simple Knowledge Organization System Reference - W3C Proposed Recommendation 15 June 2009, http://www.w3.org/TR/skos-reference.

Prud'hommeaux, Eric & Andy Seaborne. 2008. SPARQL Query Language for RDF - W3C Recommendation 15 January 2008. http://www.w3.org/TR/rdf-sparql-query/.

Vergnieux, Robert. 1985a. Karnak, Mission Permanente du CNRS. Archives documentaires, in: *Informatique et Égyptologie* 1, p. 221-223.

—. 1985b. Le fichier 'Talatat' en cours de constitution à Karnak, in: Sylvia Schoske (ed.), *Akten des vierten Internationalen Ägyptologen Kongresses, Mu̇nchen 1985* (= BSAK 1), p. 205-211.

—. 1988. Premiers exemples de résultats obtenus à l'aide du fichier informatisé sur les 'talatat' et vocabulaire de recherche, in: *Microcomputer in der Archäologie* (= WZB 37), p. 283-287.

—. 1990. La représentation des connaissances à l'aide d'unicos, in: *Informatique et Égyptologie* 7, p. 157-160.

[21] See e.g. http://www.w3.org/2007/08/pyRdfa/extract?uri=[URI to be parsed].

—. 1994. Quelques aspects d'une recherche informatisée sur un lot de talatat d'Aménophis IV, in: *Informatique et Égyptologie* 9, p. 152-155.

—. 1996. Le fac-simile électronique ou les restitutions virtuelles, in: *Informatique et Égyptologie* 10, p. 134-139.

—. 1999. *Recherches sur les monuments thébains d'Amenhotep IV à l'aide d'outils informatiques*, Genève (= Cahiers de la Société d'égyptologie 4).

VERGNIEUX, Robert & Michel GONDRAN. 1997. *Aménophis IV et les pierres du soleil : Akhénaton retrouvé*, Paris.

A Database for the Coffin Texts*

Carlos Gracia Zamacona
Madrid

1. The corpus

When conducting large empirical studies, the need for an electronic database becomes evident. This was the case for the thesis I wrote on verbs of motion in the Coffin Texts: some 200 verbs over more than 5,000 attestations,[1] and I prepared a database for that whole corpus. Until then, the only systematic study of Egyptian verbs of motion was Wente's in 1959:[2] his approach was syntactical and partly diachronic, dealing mainly with Late Egyptian, looking ahead to Coptic and occasionally backwards to Ancient and Middle Egyptian, but he did not present a closed corpus. A few cursory studies of lexical morphology,[3] lexical semantics,[4] and grammatical semantics[5] also exist, but none of these have undertaken a fully systematic survey of this topic in a closed corpus. Thus, being convinced that the functions of a linguistic unit can only be grasped by its traces (i.e. its uses in context), an exhaustive study based upon a large and closed corpus suggested itself as the best way to bring new light to these issues.[6] Such an approach relies on the characteristics of the corpus as well as on the methods and techniques used for the analysis.[7]

The Coffin Texts are the main funerary corpus written in Middle Egyptian, and thus they belong to the Middle Egyptian linguistic phase, or at least to the first sub-phase of it.[8] They are characterized by their diversity, extent, closedness and availability in full edition:

- *Diversity*. They are found on more than 200 documents (mainly rectangular wooden coffins) and composed of texts with a variety of structures — prayers, magic spells, dialogues, etc. This wide range of documents makes possible the study of dialectal variants, locally-based schools of orthography and cultural traditions, as well as diachronic change.
- *Extent*. They constitute a very large corpus: de Buck's edition (usually abbreviated as *CT*) comprises almost 3 000 pages *in folio*.

[*] I am indebted to Prof. Jean Winand for many of the theoretical and methodological concepts used for the creation and development of the database here described, and I warmly thank him for this.

[1] Gracia Zamacona 2008.

[2] Wente 1959.

[3] For example, Winand 1991; Peust 2007.

[4] For example, Depuydt 1985a & 1985b.

[5] For example, Vernus 1990; Hafemann 2001 & 2006.

[6] For the Egyptian, see mainly Schenkel 1988; Schenkel & Reineke 1998; Grunert & Hafemann 1999.

[7] See for instance Winand 1987. For a general introduction concerning Ancient Near Eastern languages, see Fronzaroli 1973.

[8] For diachronic issues on the Coffin Texts, see Vernus 1996.

- *Closedness*. They are a conceptually and functionally closed corpus. This does not mean 'canonicity' in itself, but the Coffin Texts may be considered as representing a 'phase' in the Egyptian funerary thought.[9]
- *Availability*. They are the object of a complete and critical edition made by de Buck, and then supplemented by Allen with a volume of copies of Pyramid Texts on Middle Kingdom coffins. This history of publication makes it possible to carry out analyses that are both statistically well-founded and extensively well-contrasted, and which may lead to robust conclusions.

Even though the Coffin Texts suffer from serious difficulties of interpretation, they are still an important and valuable corpus for the study of Ancient Egypt. It is true that they are frequently difficult to understand and sometimes even completely obscure, especially when they concern religious and cultural elements unknown from other sources. It is also true that the exact date of the documents cannot be ascertained at present; consequently, the border between synchrony and diachrony is also unclear — work must be conducted on documents which can be separated by up to 600 years. Finally, their genesis and textual tradition seem to have been very complex: this last field of study, complicated as it is, might eventually be unlocked through database-centered studies dealing with a great amount of data. All in all, advantages outweigh problems, and the Coffin Texts remain without doubt one of the richest sources of evidence for the study of the language, religion, and other aspects of the Egyptian civilisation.

2. THE DATABASE

The method employed is basically a contrastive one. This is a consequence of the theoretical axiom expressed above: the meaning of a linguistic unit is shown by its uses.

This method falls under the general frame of corpus or empirical linguistics,[10] and essentially consists of finding typical patterns of use and determining their extent and range. This will allow us to form questions not only about the use of linguistic features, but also about textual characteristics and varieties. Only close and systematic contact with the material combined with a contextual approach may lead to the formulation of hypotheses and their proof or disproof, because "(...) comprehensive studies of *use* cannot rely on intuition, anecdotal evidence, or small samples; they rather require empirical analysis of large databases of authentic texts, as in the corpus-based approach" [11]. Nowadays, corpus analyses are employed to deal with almost any linguistic problem.

Due to its very nature, this method requires a technique adequate and powerful enough to collect all the uses of a single item in order to distinguish differences between them. The most suitable technique seems to be an electronic database,[12] of which precision and exhaustiveness are its greatest contributions. This technique allows the quantitative analysis of data, and this is crucial for establishing general patterns. But this information must be complemented by a qualitative analysis of the data, "to examine the functional bases underlying patterns of linguistic features".[13]

The database was created with *Filemaker 4.0*, a software with an interface very intuitive for both the conception and the data entry. Beside this, *Filemaker* is very adaptable as far as the field formats,

[9] For the continuity between Pyramid Texts and Coffin Texts, see Mathieu 2004. For the Pyramid Texts copies on Middle Kingdom coffins, see Allen 2006.

[10] See, for instance, Biber, Conrad & Reppen 1998; Sampson & McCarthy 2005; Lüdeling & Kytö 2008.

[11] Biber, Conrad & Reppen 1998: 9.

[12] There was a Coffin Texts database available via the Internet (http://www.aegyptologie.uni-goettingen.de/computer/CTID/info.htm) made by P. Jürgens using *Access* software (see also Gozzoli in the current volume). On databases in Egyptian, see e.g. Winand 1990. On Coffin Texts databases and related approaches, see Gundlach & Schenkel 1970; Junge & Schenkel 1972; Hintze 1974; Schenkel 1982, 1983, 1994; Van der Plas & Borghouts 1998; Van der Molen 2005.

[13] Biber, Conrad & Reppen 1998: 139.

file connections, data ordering, and search possibilities are concerned. Later versions of *Filemaker* are differently conceived, making databases of only one file with different parts related. I have kept the older version because I consider it easier to modify depending on new analysis or textual findings. *Filemaker* files are integrated by means of records, which contain fields in which information is kept. Some of these fields are linking fields, which allow the user to relate records from different files in a logical fashion: the result is a relational database. There are two core files: MAIN VERSION (1) and, linked to it, VARIANTS (2), both in dark grey in the figure 1. The rest of the files are linked to the MAIN VERSION file and are gathered in four groups (light grey in the figure 1): three document-related files (1.1 plus two subordinates), seven word-related files (1.2 plus six subordinates), six spell-related files (1.3 plus five subordinates), and three bibliographical files (1.4 plus one subordinate and one under-subordinate, this one in soft grey). In all, the database consists of 21 files; its structure can be represented as in figure 1.

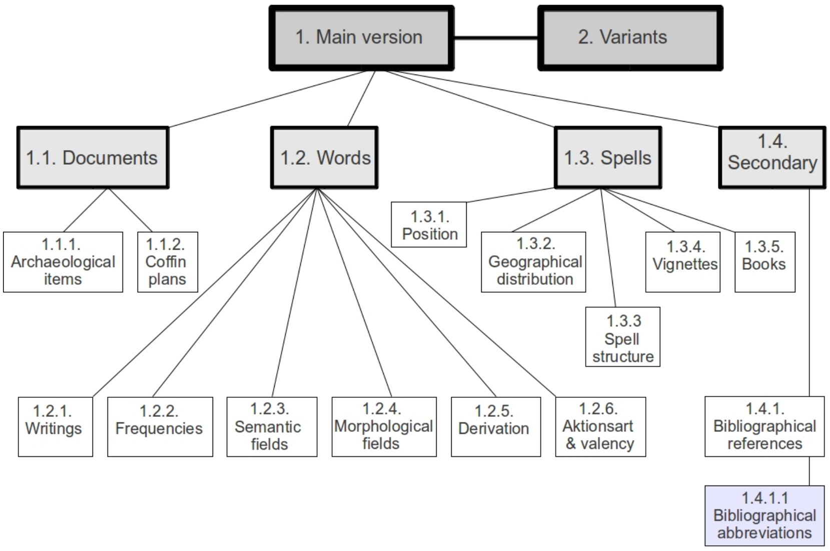


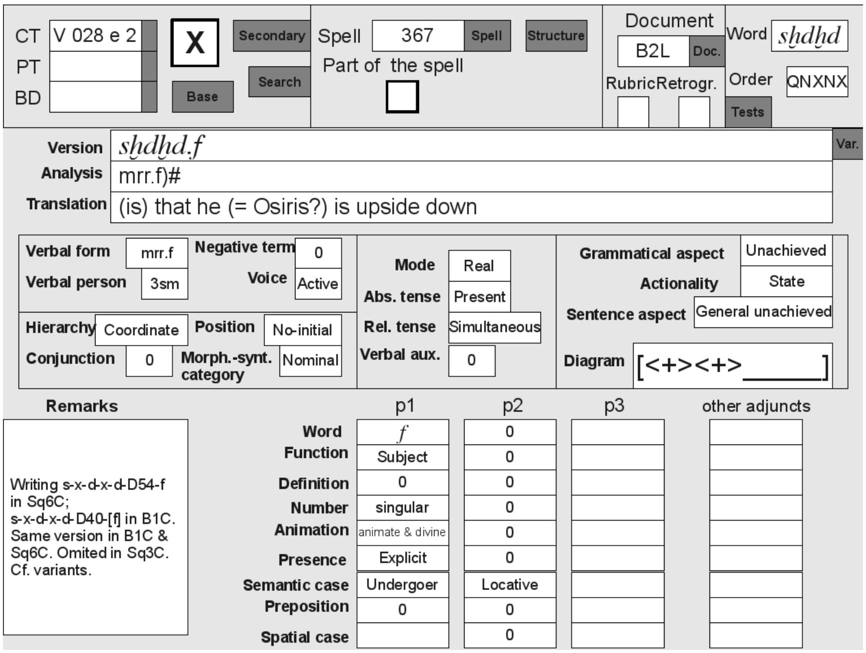
Figure 1. Structure of the database for the Coffin Texts

2.1. *Main version and variants file*

MAIN VERSION file [1] contains the passages occurring in the more complete and/or philologically more accurate document. This file consists of 27,156 records, mainly corresponding to a simple sentence each. The final scope of this file shall be to have one record for each simple sentence. The file is formed by the following fields:

- Corpus-related fields: *CT* (reference to de Buck's edition), *Spell* (of the Coffin Texts), *Document* (coffin, etc.), *Part of the spell* (referring to the spell's structure), *Rubric* (verification field for rubricated writing), *Retrograde* (verification field for retrogradate writing), *PT* (related passage in the Pyramid Texts), and *BD* (related passage in the Book of the Dead).
- Word-related fields: *Word* (terms kept in the WORDS file), and *Order* (alphabetical key for the transliteration: A = *ꜣ*, B = *ỉ*, etc. so to get the words alphabetically ordered).
- Linguistic analysis fields: *Version* (transliterated text), *Analysis* (syntactical chain), *Translation, Verbal form*, *Verbal person*, *Negative term, Voice*, *Hierarchy* (syntactical status), *Position* (syntactical position), *Conjunction* (connective term), *Morpho-syntactical category, Verbal mode, Absolute tense, Relative tense, Verbal auxiliary, Grammatical aspect, Actionality* (kind of action in context), *Sentence aspect* (complex aspectual meaning), *Diagram* (schematic

representation of the complex aspectual meaning);[14] the next fields are repeated for the first, second, and third participants, and for the adjuncts of the sentence: *Word*, *Function* (syntactical function), *Definition*, *Number*, *Animation*, *Presence*, *Semantic case, Preposition* (grammatical morpheme marking case), *Spatial case* (adjunct marking space).[15] Finally, one field is dedicated to *Remarks*.



CT V 028 e 2
PT
BD
☒
Spell 367
Part of the spell □
Document B2L
Rubric □ Retrogr. □
Word sẖdẖd
Order QNXNX

Version sẖdẖd.f
Analysis mrr.f)#
Translation (is) that he (= Osiris?) is upside down

Verbal form mrr.f
Negative term 0
Verbal person 3sm
Voice Active
Hierarchy Coordinate
Position No-initial
Conjunction 0
Morph.-synt. category Nominal
Mode Real
Abs. tense Present
Rel. tense Simultaneous
Verbal aux. 0
Grammatical aspect Unachieved
Actionality State
Sentence aspect General unachieved
Diagram [<+><+>______]

Remarks

Writing s-x-d-x-d-D54-f in Sq6C;
s-x-d-x-d-D40-[f] in B1C.
Same version in B1C & Sq6C. Omited in Sq3C.
Cf. variants.

	p1	p2	p3	other adjuncts
Word	*f*	0		
Function	Subject	0		
Definition	0	0		
Number	singular	0		
Animation	animate & divine	0		
Presence	Explicit	0		
Semantic case	Undergoer	Locative		
Preposition	0	0		
Spatial case		0		

Figure 2. Main version file

The linking fields of the MAIN VERSION file are as follows: *CT* to SECONDARY, and VARIANTS; *Document* to DOCUMENTS; *Order* to WORDS; and *Spell* to SPELLS.[16]

VARIANTS file [2] contains the textual variations of the main version (i.e. the more complete and/or philologically more accurate document). It has the same fields (and linking fields) as MAIN VERSION file (1). The number of records is 6,512.

2.2. *Documents file*

DOCUMENTS file [1.1] stocks information on the documents bearing Coffin Texts (see figure 3). These are its fields:

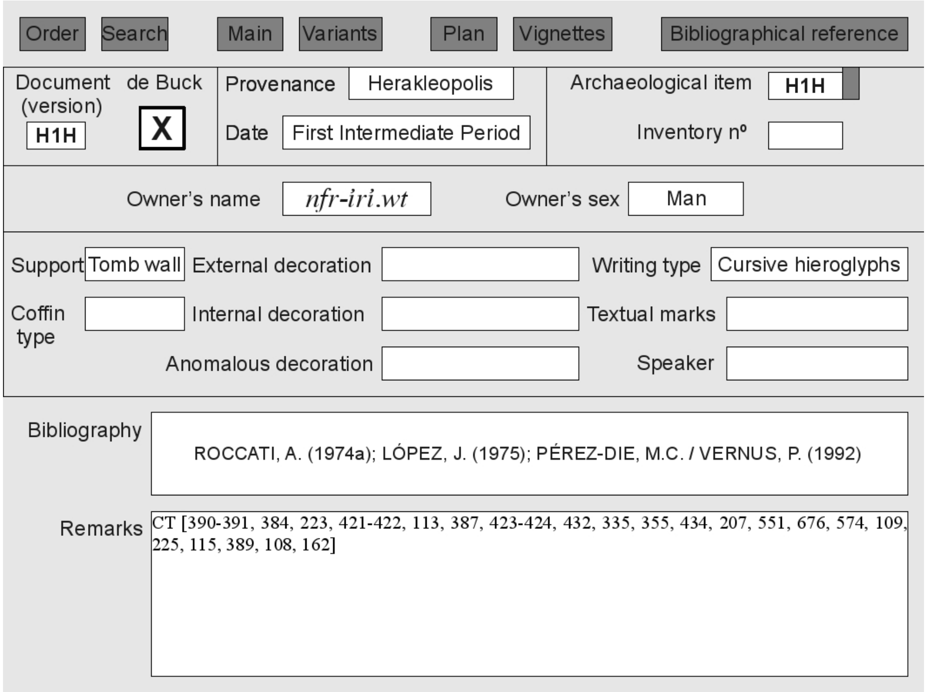
- One field called *Document (version)*: textual version of a given document — a document can have more than one version of a given text.
- Corpus-related field *de Buck*: verification field on the occurrence of a given document in de Buck's edition.
- Document-related fields: *Provenance*, *Date*, *Archaeological item* (archaeological item's name), *Inventory number*, *Support* (coffin, tomb wall...), *Coffin type* (internal, middle or external), *External decoration*, *Internal decoration*, *Anomalous decoration*,[17] *Owner's name*, *Owner's sex*, *Writing type*, *Textual marks*, *Speaker* (1st/3rd person), *Bibliography*, and *Remarks*.

[14] According to Winand 2006: 121.

[15] For the concept, see Gracia Zamacona 2010b.

[16] In this and the following figures, buttons activating links between files are represented by dark rectangles.

[17] The last three fields pertaining to decoration are defined according to Willems 1988.



Document (version): H1H
de Buck: ☒
Provenance: Herakleopolis
Date: First Intermediate Period
Archaeological item: H1H
Inventory nº:

Owner's name: *nfr-iri.wt*
Owner's sex: Man

Support: Tomb wall
Coffin type:
External decoration:
Internal decoration:
Anomalous decoration:
Writing type: Cursive hieroglyphs
Textual marks:
Speaker:

Bibliography: ROCCATI, A. (1974a); LÓPEZ, J. (1975); PÉREZ-DIE, M.C. / VERNUS, P. (1992)

Remarks: CT [390-391, 384, 223, 421-422, 113, 387, 423-424, 432, 335, 355, 434, 207, 551, 676, 574, 109, 225, 115, 389, 108, 162]

Figure 3. Documents file

In total, the documents file contains 330 records. The linking fields are *Document* to MAIN VERSION (see 2.1), and *Archaeological item* to ARCHAEOLOGICAL ITEM and COFFIN PLAN:

– ARCHAEOLOGICAL ITEM file [1.1.1] holds archaeological information, mainly graphical, about the documents. Its fields are: *Archaeological item* (archaeological item's name), *Item's part* (lid, front side...), *Image*, *Bibliography*, and *Remarks*.
– COFFIN PLAN file [1.1.2] provides a schematic representation of the graphical arrangement of texts and vignettes in a given document. Its fields are: *Document* (document's version), *Archaeological item* (archaeological item's name) *Bibliography*, and *Image*.

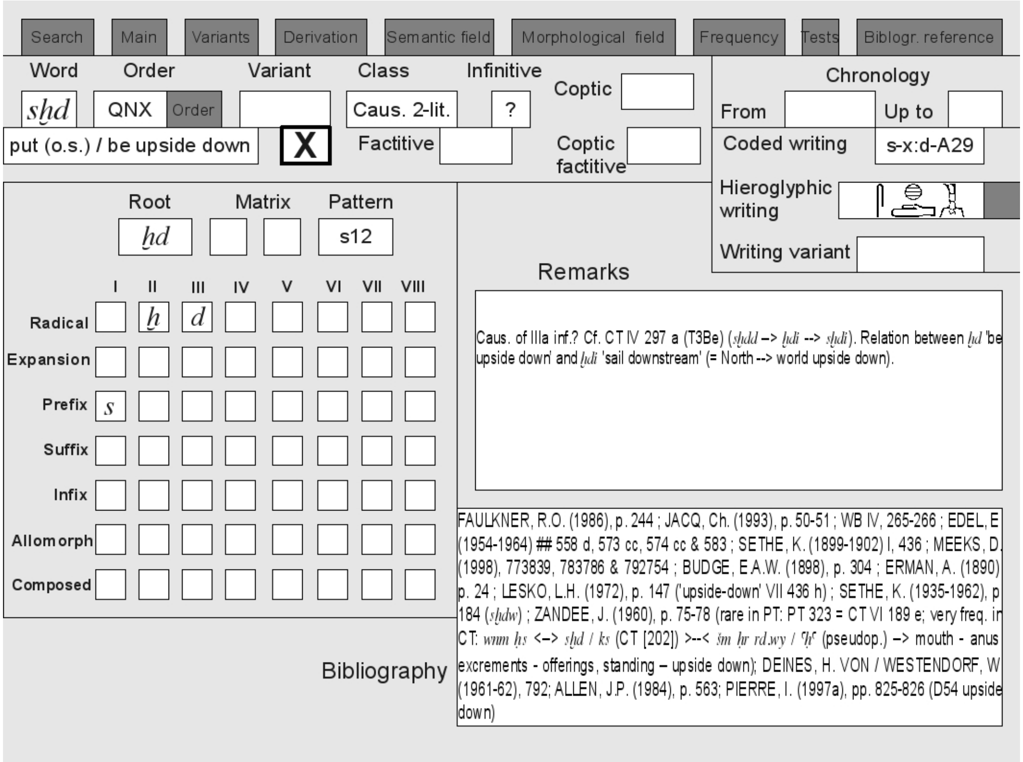
2.3. *Words file*

The words in the Coffin Texts are listed in this file [1.2]. Only verbs (mainly verbs of motion) have been currently introduced: 724 records. The fields are:

– Word-related fields: *Word* (transliteration), *Order* (alphabetical key), *Variant* (word variant), *Class* (uniliteral, etc.), *Infinitive* (infinitive type), *Meanings*, *Coptic* (Coptic descendant), *Factitive*, *Coptic factitive*, *Coded writing* (coded standard writing)[18], *Hieroglyphic writing* (hieroglyphic standard writing), *Writing variant* (main hieroglyphic variant), *From* (first attested in Egyptian), *Up to* (last attested in Egyptian), and *Bibliography*.
– Corpus-related field: *CT* (verification field on the occurrence of a given word in the Coffin Texts).
– Fields for analysis of lexical morphology and semantics: *Root*, *Matrix 1-2* (matrix elements),[19] *Pattern* (root pattern: for instance, 123 for a triliteral root); the next fields are repeated for the theoretical positions I-VIII of phonemes: *Radical (Rd)* (radical element), *Expansion (Ex)* (expanded element), *Prefix (Pf)* (prefixed element), *Suffix (Sf)* (suffixed element), *Infix (If)* (infixed element), *Allomorph (All)* (allomorphic element), and *Composed (Co)* (composed element). A field for *Remarks* is also provided.

[18] I.e. the more common writing of a given word in the Coffin Texts.

[19] According to Bohas 1997.



Word	*sḫd*
Order	QNX
Variant	
Class	Caus. 2-lit.
Infinitive	?
Coptic	
	put (o.s.) / be upside down
	☒
Factitive	
Coptic factitive	
Chronology From	
Up to	
Coded writing	s-x:d-A29
Hieroglyphic writing	
Writing variant	
Root	*ḫd*
Matrix	
Pattern	s12

	I	II	III	IV	V	VI	VII	VIII
Radical		*ḫ*	*d*					
Expansion								
Prefix	*s*							
Suffix								
Infix								
Allomorph								
Composed								

Remarks

Caus. of IIIa inf.? Cf. CT IV 297 a (T3Be) (*sḫdd* –> *ḫdi* --> *sḫdi*). Relation between *ḫd* 'be upside down' and *ḫdi* 'sail downstream' (= North --> world upside down).

Bibliography

FAULKNER, R.O. (1986), p. 244 ; JACQ, Ch. (1993), p. 50-51 ; WB IV, 265-266 ; EDEL, E (1954-1964) ## 558 d, 573 cc, 574 cc & 583 ; SETHE, K. (1899-1902) I, 436 ; MEEKS, D. (1998), 773839, 783786 & 792754 ; BUDGE, E.A.W. (1898), p. 304 ; ERMAN, A. (1890) p. 24 ; LESKO, L.H. (1972), p. 147 ('upside-down' VII 436 h) ; SETHE, K. (1935-1962), p 184 (*sḫdw*) ; ZANDEE, J. (1960), p. 75-78 (rare in PT: PT 323 = CT VI 189 e; very freq. in CT: *wnm ḥs* <–> *sḫd / ks* (CT [202]) >--< *šm ḥr rd.wy / ꜥḥꜥ* (pseudop.) –> mouth - anus excrements - offerings, standing – upside down); DEINES, H. VON / WESTENDORF, W (1961-62), 792; ALLEN, J.P. (1984), p. 563; PIERRE, I. (1997a), pp. 825-826 (D54 upside down)

Figure 4. Words file

Order is the linking field to Main version and Variants, but also to Writings, Frequencies, Semantic field, Morphological field, Derivation, and Aktionsart & valency.

2.3.1. Writings file

Information pertaining to the ways in which words are written is to be compiled in this file [1.2.1]. The study of different word “spellings” is crucial not only to know the use of the hieroglyphic signs (a matter of study in itself), but also for the study of lexical semantics. The fields are:

- Word-related fields: *Word* (words in transliteration), *Order* (alphabetical key), and *Coded writing*.
- Corpus-related field: *CT* (main version).
- Linguistic-related fields: Grammatical category (noun, verb, etc.), Verbal form, Verbal person, Status (absolute, construct, pronominal), Pronoun type, Particle type, and Preposition/Conjunction type.

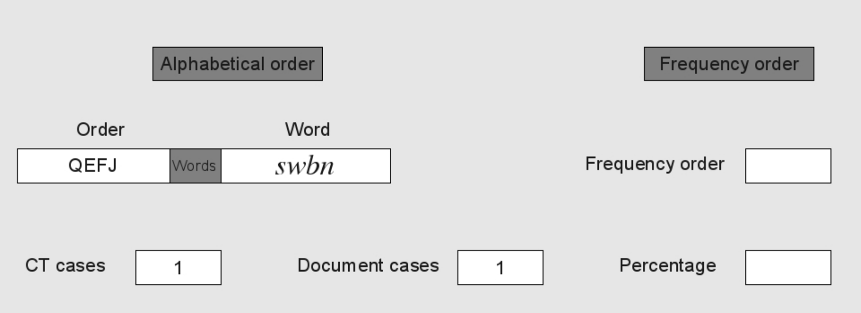
2.3.2. Frequencies file

The purpose of this file [1.2.2] is to record word frequencies, as a first step to studying word uses. For instance, the recording of word frequencies can be highly relevant for the study of relationships between semantically marked words and morpho-syntactically marked constructions.[20] This file contains the following fields:

- Word-related fields: *Word* (words in transliteration), and *Order* (alphabetical key).
- Frequency-related fields: *CT cases* (number of occurrences of a word in the Coffin Texts), *Document cases* (number of documents in which a word appears), *Percentage* (percentage on

[20] This seems, for example, to be the case of the alternation ‘that’ / Ø for object clauses in English; see Rissanen 2005: 145.

the total number of word frequencies), and *Frequency order* (Order number of a word frequency).



Order: QEFJ

Word: *swbn*

Frequency order:

CT cases: 1

Document cases: 1

Percentage:

Figure 5. Frequencies file

2.3.3. Semantic field file

This file [1.2.3] stores information on the meaning relationships of each word. Its fields are:

- Word-related fields: *Word* (words in transliteration), and *Order* (alphabetical key).
- Fields related to the semantic field: *Semantic field* (Semantic field's label), *Scheme* (Semantic field's conceptual scheme, below the previous field), *Synonyms*, *Antonyms*, *Hyperonyms*, *Hyponyms*, *Co-hyponyms*, *Meronyms*, and *Homonyms*.

2.3.4. Morphological field file

This file [1.2.4] gathers information about the formal relationships of words that are semantically related to each other. Its fields are:

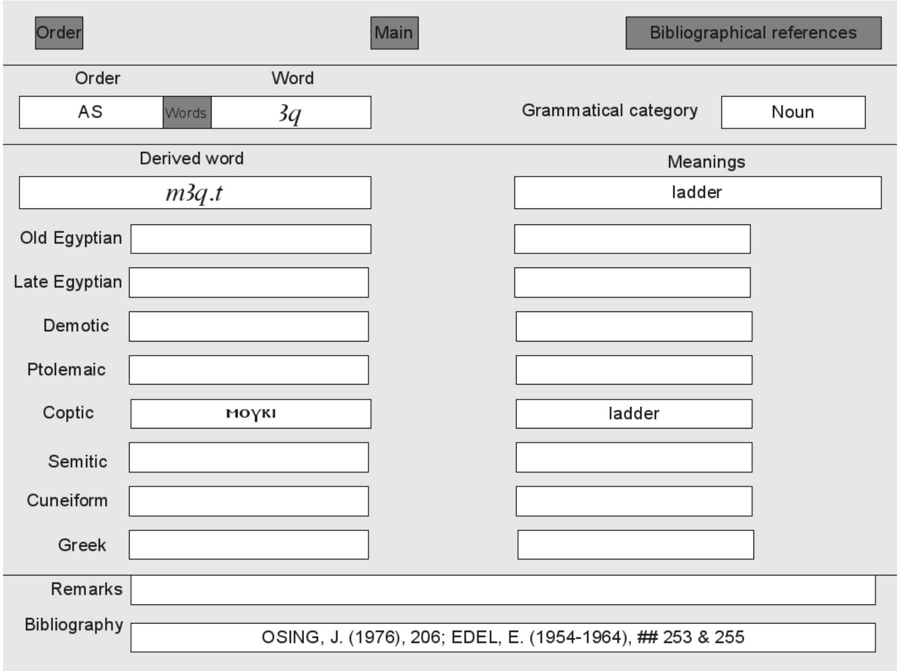
- Word-related fields: *Word* (words in transliteration), and *Order* (alphabetical key).
- Fields related to the morphological field: *Morphological field* (morphological field's label), *Scheme* (morphological field's conceptual scheme, below the previous one), *Related words* (possible morphologically related words), and *Composition* (related composed words).

2.3.5. Derivation file

This file [1.2.5] has information about derivative processes for each word. Such information would be of interest for lexical morphology, mainly from a diachronic point of view. Its fields are of two kinds:

- Word-related fields: *Word* (transliteration), *Order* (alphabetical key), *Grammatical category* (verb, noun, etc.), *Derived word* (Middle Egyptian derived word, transliterated), and *Meaning* (meaning of the Middle Egyptian derived word).[21]
- Secondary fields: *Bibliography* (concerning the word studied in each record), and *Remarks* (free-content field).

[21] The last two fields are repeated for the other phases of Egyptian — Old Egyptian, Late Egyptian, Demotic, Ptolemaic (for the sake of exhaustiveness, as it is not properly a linguistic phase) and Coptic, including foreign equivalents Semitic, Cuneiform and Greek.



Order	AS
Word	*ȝq*
Grammatical category	Noun
Derived word	*mȝq.t*
Meanings	ladder
Coptic	ⲙⲟⲩⲕⲓ — ladder
Bibliography	OSING, J. (1976), 206; EDEL, E. (1954-1964), ## 253 & 255

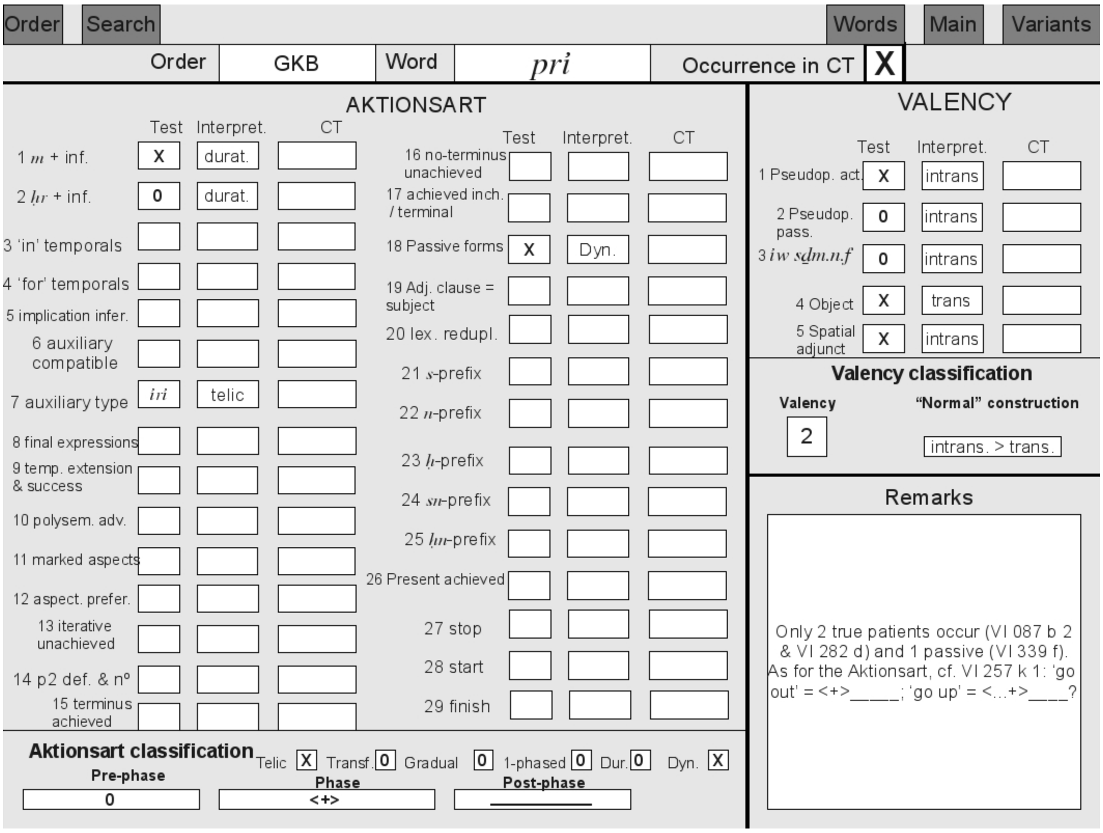
Figure 6. Derivation file

2.3.6. Aktionsart and valency file

Tests to determine the valency and Aktionsart structure of words are included in this file [1.2.6]. Current number of records is 241. This file consists of the following three kinds of fields:

- Word-related fields: *Word* (words in transliteration), and *Order* (alphabetical key).
- Text-related field: *Occurrence in CT* (verification field about the presence of the word in the Coffin Texts).
- Linguistic fields: *Aktionsart 1-29* (29 Aktionsart tests, each one with three fields: verification, interpretation and occurrence in Coffin Texts), *Valency 1-5* (5 valency tests, each one with three fields: verification, interpretation and occurrence in Coffin Texts), *Telic* (verification field about telicity), *Transformative* (verification field about transformativity), *Gradual* (verification field about graduality), *One-phased* (verification field about phases), *Durative* (verification field about duration), *Dynamic* (verification field about dynamicity), *Pre-phase* (schematic representation of the prephase or 0), *Phase* (schematic representation of the phase), *Post-phase* (schematic representation of the postphase or 0),[22] *Valency* (valency classification: 2 — unitransitive —, etc.), *"Normal" construction* (transitive, transitive > intransitive, etc.).

[22] The last three fields concerning phase are defined according to Winand 2006: 121.



Order GKB Word *pri* Occurrence in CT ☒

AKTIONSART

	Test	Interpret.	CT
1 *m* + inf.	X	durat.	
2 *ḥr* + inf.	0	durat.	
3 'in' temporals			
4 'for' temporals			
5 implication infer.			
6 auxiliary compatible			
7 auxiliary type	*iri*	telic	
8 final expressions			
9 temp. extension & success			
10 polysem. adv.			
11 marked aspects			
12 aspect. prefer.			
13 iterative unachieved			
14 p2 def. & nº			
15 terminus achieved			

	Test	Interpret.	CT
16 no-terminus unachieved			
17 achieved inch. / terminal			
18 Passive forms	X	Dyn.	
19 Adj. clause = subject			
20 lex. redupl.			
21 *s*-prefix			
22 *n*-prefix			
23 *ḥ*-prefix			
24 *sn*-prefix			
25 *ḥn*-prefix			
26 Present achieved			
27 stop			
28 start			
29 finish			

Aktionsart classification Telic ☒ Transf. 0 Gradual 0 1-phased 0 Dur. 0 Dyn. ☒

Pre-phase	Phase	Post-phase
0	<+>	________

VALENCY

	Test	Interpret.	CT
1 Pseudop. act.	X	intrans	
2 Pseudop. pass.	0	intrans	
3 *iw sḏm.n.f*	0	intrans	
4 Object	X	trans	
5 Spatial adjunct	X	intrans	

Valency classification

Valency	"Normal" construction
2	intrans. > trans.

Remarks

Only 2 true patients occur (VI 087 b 2 & VI 282 d) and 1 passive (VI 339 f). As for the Aktionsart, cf. VI 257 k 1: 'go out' = <+>_____; 'go up' = <...+>____?

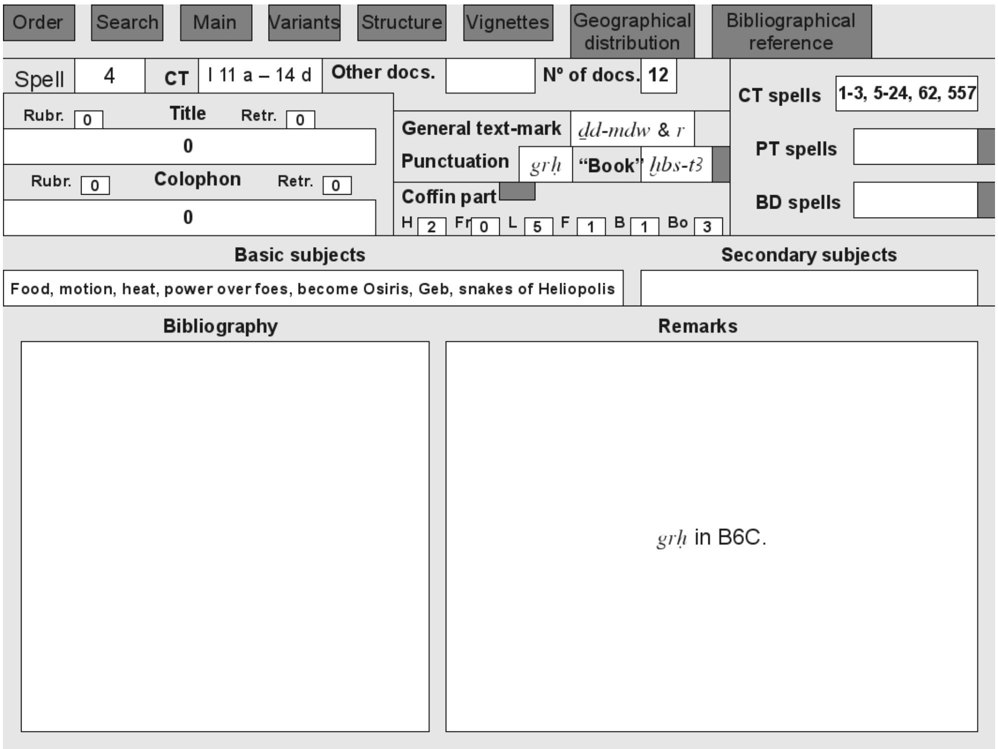
Figure 7. Aktionsart and valency file

2.4. *Spell File*

This file [1.3] has 1,185 records, the number of spells in the Coffin Texts according to de Buck's edition. This is a key file to analyse the corpus' textual structure. It is made up by six types of fields:

- Corpus-related fields: *Spell* (Coffin Texts spell number), and *CT* (spell textual extension in the Coffin Texts).
- Document-related fields: *Other documents* (new documents containing a given spell), *Number of documents* (number of documents containing a given spell), *Head (H)* (number of documents in which a given spell is written on the "head"-side), *Front (Fr)* (the same on the "front"-side), *Lid (L)* (the same on the lid), *Foot (F)* (the same on the "foot"-side), *Back (B)* (the same on the "back"-side), and *Bottom (Bo)* (the same on the bottom).[23]
- Intertextual fields: *CT spells* (related spells in the Coffin Texts), *PT spells* (the same in the Pyramid Texts), and *BD spells* (the same in the Book of the Dead).
- Spell structure-related fields: *Title* (spell title), *Colophon* (spell colophon), *Rubric* (verification field on rubricated writing for *Title* and *Colophon* fields), *Retrogradate* (the same on retrogradate writing), *Basic subjects, Secondary subjects*, *General text-mark* (*r* 'spell', *mḏꜣ.t* 'book', etc.), and *Punctuation* (*grḥ*-sign, etc.).
- "Book"-related field: *"Book"* (name of the "book" a given spell belongs to).
- Secondary fields: *Bibliography* (on the spell studied in each record), and *Remarks* (free-content field).

[23] Nomenclature according to Willems 1988.



Spell 4 CT I 11 a – 14 d Other docs. N° of docs. 12

Rubr. 0 Title Retr. 0

0

Rubr. 0 Colophon Retr. 0

0

General text-mark *ḏd-mdw* & *r*

Punctuation *grḥ* "Book" *ḫbs-tꜣ*

Coffin part

H 2 Fr 0 L 5 F 1 B 1 Bo 3

CT spells 1-3, 5-24, 62, 557

PT spells

BD spells

Basic subjects

Food, motion, heat, power over foes, become Osiris, Geb, snakes of Heliopolis

Secondary subjects

Bibliography

Remarks

grḥ in B6C.

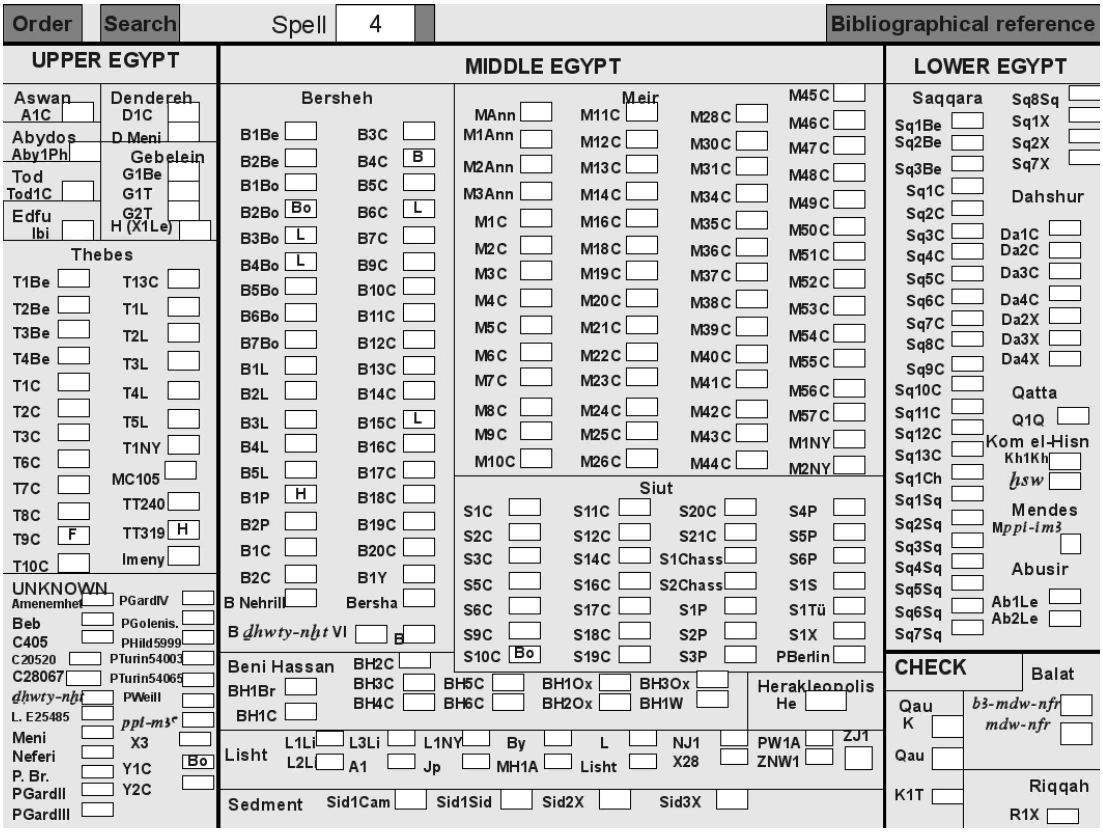
Figure 8. Spell file

Spell is the linking field to Main version as well as to Position, Geographical distribution, Spell structure, Vignettes, and Book.

2.4.1. Position file

In this file [1.3.1], each document has a field in which the position of each spell is marked (256 fields in all): head (H), foot (F), etc. Summary results are presented in the fields assigned in the SPELLS file. There is also a field called *Spell*, which is the linking field to SPELLS. In total, 1 185 records integrate this file. The fields are grouped by geographical zone (Upper Egypt, Middle Egypt, Lower Egypt, etc.), by locality inside of each zone (e.g. Bersheh, Meir, Siut, etc. in Middle Egypt), and archaeological item (for example, Meir coffins: M1C, M2C, etc.).

This file, and the following one (Geographical Distribution file), are crucial to make a sound study on why a certain spell appears on a given side (or several sides) of a coffin, and which kind of factors may be involved in this, for instance, local traditions.


Order
Search
Bibliographical reference
UPPER EGYPT
MIDDLE EGYPT
LOWER EGYPT
Aswan A1C
Dendereh D1C D Meni
Abydos Aby1Ph
Gebelein G1Be G1T G2T H (X1Le)
Tod Tod1C
Edfu Ibi
Thebes T1Be T2Be T3Be T4Be T1C T2C T3C T6C T7C T8C T9C T10C T13C T1L T2L T3L T4L T5L T1NY MC105 TT240 TT319 Imeny
UNKNOWN Amenemhet Beb C405 C20520 C28067 ḏḥwty-nḫt L. E25485 Meni Neferi P. Br. PGardII PGardIII PGardIV PGolenis. PHild5999 PTurin54003 PTurin54065 PWeill ppi-mꜣꜥ X3 Y1C Y2C
Bersheh B1Be B2Be B1Bo B2Bo B3Bo B4Bo B5Bo B6Bo B7Bo B1L B2L B3L B4L B5L B1P B2P B1C B2C B Nehrill B ḏḥwty-nḫt VI B3C B4C B5C B6C B7C B9C B10C B11C B12C B13C B14C B15C B16C B17C B18C B19C B20C B1Y Bersha B
Meir MAnn M1Ann M2Ann M3Ann M1C M2C M3C M4C M5C M6C M7C M8C M9C M10C M11C M12C M13C M14C M16C M18C M19C M20C M21C M22C M23C M24C M25C M26C M28C M30C M31C M34C M35C M36C M37C M38C M39C M40C M41C M42C M43C M44C M45C M46C M47C M48C M49C M50C M51C M52C M53C M54C M55C M56C M57C M1NY M2NY
Siut S1C S2C S3C S5C S6C S9C S10C S11C S12C S13C S16C S17C S18C S19C S20C S21C S1Chass S2Chass S1P S2P S3P S4P S5P S6P S1S S1Tü S1X PBerlin
Beni Hassan BH1Br BH1C BH2C BH3C BH4C BH5C BH6C BH1Ox BH2Ox BH3Ox BH1W
Herakleopolis He
Lisht L1Li L2Li L3Li A1 L1NY Jp By MH1A L Lisht NJ1 X28 PW1A ZNW1 ZJ1
Sedment Sid1Cam Sid1Sid Sid2X Sid3X
Saqqara Sq1Be Sq2Be Sq3Be Sq1C Sq2C Sq3C Sq4C Sq5C Sq6C Sq7C Sq8C Sq9C Sq10C Sq11C Sq12C Sq13C Sq1Ch Sq1Sq Sq2Sq Sq3Sq Sq4Sq Sq5Sq Sq6Sq Sq7Sq Sq8Sq Sq1X Sq2X Sq7X
Dahshur Da1C Da2C Da3C Da4C Da2X Da3X Da4X
Qatta Q1Q
Kom el-Hisn Kh1Kh ḫsw
Mendes Mppi-im3
Abusir Ab1Le Ab2Le
CHECK
Balat bꜣ-mdw-nfr mdw-nfr
Qau K Qau K1T
Riqqah R1X

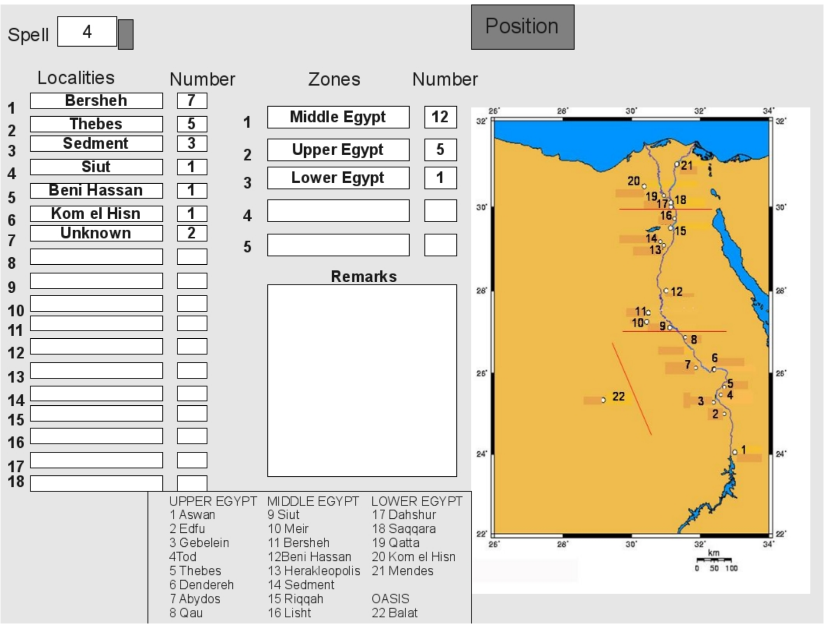
Spell 4

B2Bo Bo
B3Bo L
B4Bo L
B1P H
B4C B
B6C L
B15C L
T9C F
TT319 H
Y1C Bo
S10C Bo

Figure 9. Position file

2.4.2. Geographical distribution file

Apart of a field called *Spell* with the Coffin Texts spell number and another for *Remarks*, the fields in this file [1.3.2] are geographically and quantitatively arranged by localities (*1-18: locality* and *1-18: number* [number of occurrence of each spell in each locality]), and by zones (*1-5: zone* and *1-5: number* [number of occurrence of each spell in each zone]). A field called *Map* (map displaying the geographical distribution of a given spell) is also provided.



Spell 4

	Localities	Number
1	Bersheh	7
2	Thebes	5
3	Sedment	3
4	Siut	1
5	Beni Hassan	1
6	Kom el Hisn	1
7	Unknown	2
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		

	Zones	Number
1	Middle Egypt	12
2	Upper Egypt	5
3	Lower Egypt	1
4		
5		

Remarks

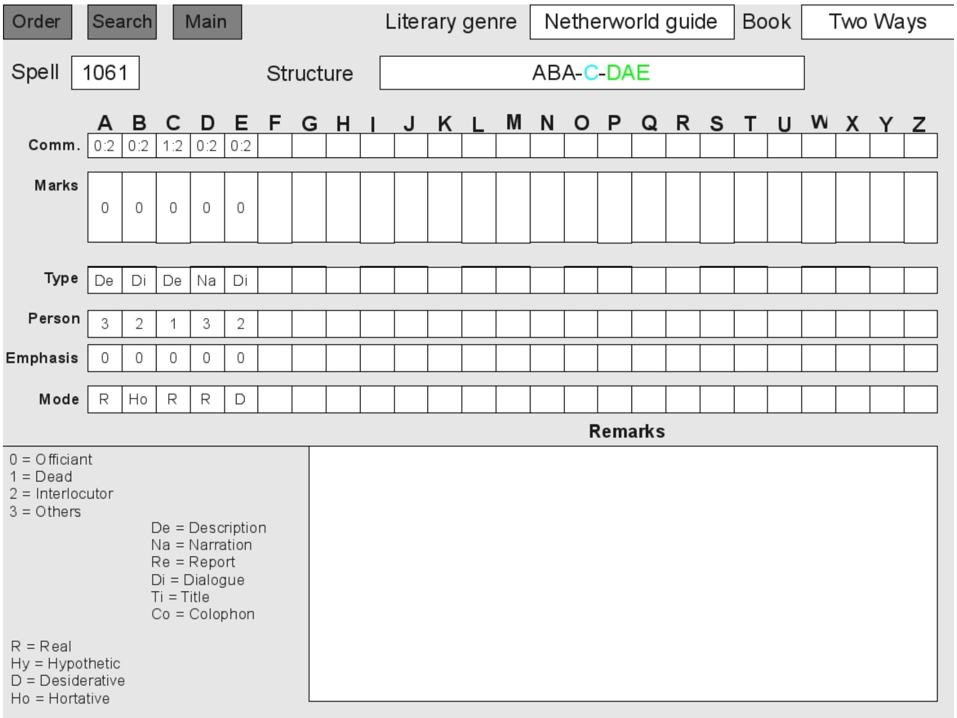
UPPER EGYPT	MIDDLE EGYPT	LOWER EGYPT
1 Aswan	9 Siut	17 Dahshur
2 Edfu	10 Meir	18 Saqqara
3 Gebelein	11 Bersheh	19 Qatta
4Tod	12Beni Hassan	20 Kom el Hisn
5 Thebes	13 Herakleopolis	21 Mendes
6 Dendereh	14 Sedment	
7 Abydos	15 Riqqah	OASIS
8 Qau	16 Lisht	22 Balat

Figure 10. Geographical distribution file

2.4.3. Spell structure file

This file [1.3.3] has information about the textual structure of each spell. This is a file that is key to studying literary genres and textual typology in the Coffin Texts. It may also be important in studying larger textual unities, i.e. the so-called "books". The fields are grouped into three main classes:

- Corpus-related field: *Spell* (Coffin Texts spell number).
- Fields related to the spell structure: *Structure* (schematic line of letters representing the main structure of a spell: e.g. ABBA), *Literary genre* (for example, netherworld guide, dramatic text). The next fields repeat from A to Z the possible parts: *Communication* (communicative agents: 0 = unspecified (usually a celebrant), 1 = dead, etc.),[24] *Marks* (textual marks, e.g. *ky-ḏd* 'gloss', etc.), *Type* (textual type: narration, dialogue, etc.), *Person* (first person, etc.), *Emphasis* (verification field for emphasis), and *Mode* (real, hypothetical, etc.).
- "Book"-related field: *"Book"* (name of the "book").



Literary genre: Netherworld guide
Book: Two Ways
Spell: 1061
Structure: ABA-C-DAE

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	W	X	Y	Z
Comm.	0:2	0:2	1:2	0:2	0:2																				
Marks	0	0	0	0	0																				
Type	De	Di	De	Na	Di																				
Person	3	2	1	3	2																				
Emphasis	0	0	0	0	0																				
Mode	R	Ho	R	R	D																				

0 = Officiant
1 = Dead
2 = Interlocutor
3 = Others

De = Description
Na = Narration
Re = Report
Di = Dialogue
Ti = Title
Co = Colophon

R = Real
Hy = Hypothetic
D = Desiderative
Ho = Hortative

Figure 11. Spell structure file

2.4.4. Vignettes file

This file [1.3.4] provides information about each spell vignette, a rare issue in the Coffin Texts. The fields of this file are:

- Corpus-related field: *Spell* (CT spell number).
- Document-related fields: *Archaeological object* (object with CT), *Document* (textual version on a coffin, etc.),[25] and *Image* (vignette picture / drawing).
- Secondary fields: Bibliographical reference, and Remarks.

[24] See Assmann 1990: 6.

[25] An object can keep more than one version of a given text.

2.4.5. Book file

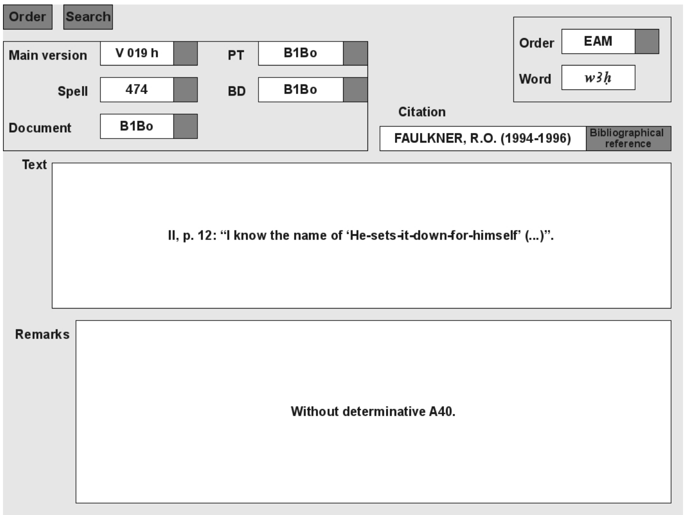
This file [1.3.5] has information about possible textual units superior to individual spells: the so-called "books". The file has 39 records at present. It consists of the following three kinds of fields:

- "Book"-related fields: *"Book"* (new title of the "book"), *CT spells* (CT spell numbers belonging to a given "book"), *Title* (original title of the "book"), and *Colophon* (colophon of the "book").
- Intertextual fields: *PT* (related spells in the Pyramid Texts), and *BD* (related spells in the Book of the Dead).
- Secondary fields: *Bibliographical reference* (to the "book"), and *Remarks*.

2.5. *Secondary file*

This file [1.4] offers information from secondary literature on each record of the MAIN VERSION file [1].[26] The fields of this file are:

- Corpus-related fields: *Main version* (main textual version), *CT spell* (CT spell number), and *Document* (document name).
- Intertextual fields: *PT* (related spells in the Pyramid Texts), and *BD* (related spells in the Book of the Dead).
- Word-related fields: *Order* and *Word*.
- Secondary fields: *Citation* (bibliographical reference quoted), *Text* (quotation), and *Remarks*.



Main version V 019 h
Spell 474
Document B1Bo
PT B1Bo
BD B1Bo
Order EAM
Word *w3ḥ*
Citation FAULKNER, R.O. (1994-1996)
Text II, p. 12: "I know the name of 'He-sets-it-down-for-himself' (...)".
Remarks Without determinative A40.

Figure 12. Secondary file

The linking fields are *Main version* to MAIN VERSION, *Spell* to SPELLS, *Document* to DOCUMENTS, and *Citation* to BIBLIOGRAPHICAL REFERENCE.

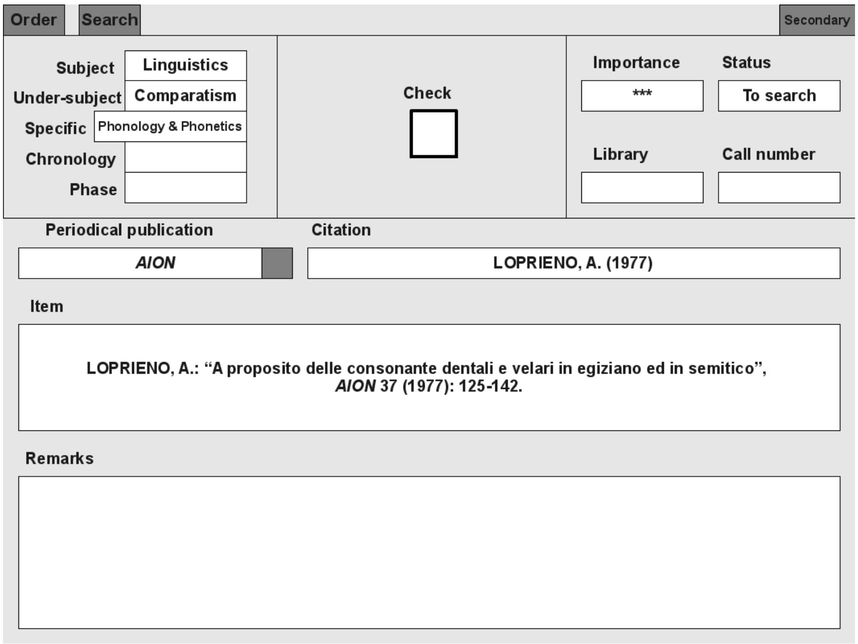
2.5.1. *Bibliographical reference file*

In this file [1.4.1] the bibliographical information is displayed. The file attains at present 6,748 records, mainly in Egyptology and Linguistics. Fields fall into three groups:

- Logical fields: *Subject* (general field of knowledge, e.g. 'linguistics'), *Under-subject* (field further specifying the *subject* field, e.g. 'semantics'), *Specific* (specific subjects, e.g. 'Aktionsart'), *Chronology* (general temporal period in Egypt), and *Phase* (linguistic phase of Egyptian).

[26] Compare Grieshammer 1974.

- Logistical fields: *Importance* (interest of the reference), *Status* (e.g. 'photocopied', 'to be read', etc.), *Library*, *Call number*, and *Verification* (verification field for records to be checked).
- Bibliographical fields: *Citation* (compact bibliographical reference), *Periodical publication* (abbreviated name of periodical, series, etc.), and *Item* (full bibliographical reference).



Subject: Linguistics
Under-subject: Comparatism
Specific: Phonology & Phonetics
Chronology:
Phase:
Check: □
Importance: ***
Status: To search
Library:
Call number:
Periodical publication: *AION*
Citation: LOPRIENO, A. (1977)
Item: LOPRIENO, A.: "A proposito delle consonante dentali e velari in egiziano ed in semitico", *AION* 37 (1977): 125-142.
Remarks:

Figure 13. Bibliographical reference file

The linking fields are *Citation* to SECONDARY, and *Periodical* to BIBLIOGRAPHICAL ABBREVIATION.

BIBLIOGRAPHICAL ABBREVIATION file [1.4.1.1] contains information on periodicals, series, collective books, homage volumes, conference proceedings, etc. It has 1 272 records at present.

3. PRESENT RESULTS AND POSSIBLE DEVELOPMENTS

3.1. *Present results*

A database like this is a research tool fundamental to providing answers to questions otherwise impossible to reach. It makes possible complementary approaches to meaning on quantitative as well as qualitative levels. In order to give a better idea of the analytical possibilities this tool could offer, an example of each kind of approach is outlined below:

- *Quantitative*. The general behaviour of prepositions as space markers and their relationship to verbs of motion in the Coffin Texts display almost no sign of 'government'. In fact, what motivates the selection of a given preposition is the particular instance of spatial expression intended and the semantic restrictions on the kind of entity introduced by a given preposition in a specific context.[27]
- *Qualitative*. The behaviour of the verb *ʿq* 'enter' with prepositions could seem aberrant at first sight,[28] but it is in fact a crucial illustration of the interaction between its valency and Aktionsart, and is semantically 'logical', so to speak. It constitutes an excellent case of how a human language works. For this case, all the instances of *ʿq* in the Coffin Texts were checked to be sure that no other factors (such as the spatial properties of 'doors', intimately connected to this verb) have any effect on its behaviour: the fact is that in the Coffin Texts, the verb *ʿq*

[27] See now Gracia Zamacona 2010a & 2010b.

[28] Winand 2006: n. 84; Nyord 2010: 34-35.

'enter' is followed by the preposition *m* 'in' or *r* 'to' to express the illative dependent on the status of the spatial adjunct as 'substance' or 'object' respectively.[29]

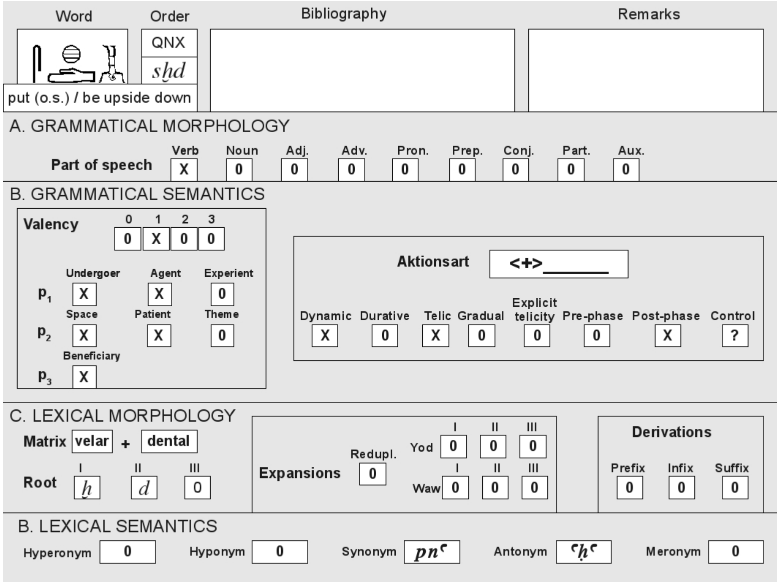
3.2. *Possible developments*

Databases like this one are intended to constitute an instrument for future research not only in Egyptology, but also in other academic fields like Linguistics or the History of Religions because of the very nature of the Coffin Texts: one of the oldest and largest funerary corpora of human civilisation written in one of the longest-attested languages in the world.

A full development of this database is certainly desired and needed in order to make it completely operative and successful. Such a development would imply: first, completing data entry for all files; second, adapting the database to a general user, which involves matching every record of the MAIN VERSION file with every simple sentence in the Coffin Texts; third, enlarging the database with the aim of creating an interrelated database of all Egyptian funerary texts.

In order to accomplish these goals, one or more research teams should be created, especially with regard to the third objective. I would like this article to be considered as an invitation to a fruitful collaboration more than an advertisement of individual desire. Indeed, we can find without doubt an inspiring model in the research group *Ramsès*[30] under the direction of St. Polis and Prof. Winand of the University of Liège.

To conclude, a specific example of this kind of development would be the creation of a linguistic dictionary of the Coffin Texts, using here the term 'linguistic' as opposed to 'encyclopaedic', as dictionnaries used to be.[31] A first draft of it could look something like the following:



Word	Order	Bibliography	Remarks
put (o.s.) / be upside down	QNX *sẖd*		

A. GRAMMATICAL MORPHOLOGY

	Verb	Noun	Adj.	Adv.	Pron.	Prep.	Conj.	Part.	Aux.
Part of speech	X	0	0	0	0	0	0	0	0

B. GRAMMATICAL SEMANTICS

	0	1	2	3
Valency	0	X	0	0

	Undergoer	Agent	Experient
p₁	X	X	0
	Space	Patient	Theme
p₂	X	X	0
	Beneficiary		
p₃	X		

Aktionsart <+>______

Dynamic	Durative	Telic	Gradual	Explicit telicity	Pre-phase	Post-phase	Control
X	0	X	0	0	0	X	?

C. LEXICAL MORPHOLOGY

Matrix velar + dental

	I	II	III
Root	*ẖ*	*d*	0

Expansions

		I	II	III
Redupl. 0	Yod	0	0	0
	Waw	0	0	0

Derivations

Prefix	Infix	Suffix
0	0	0

B. LEXICAL SEMANTICS

Hyperonym	Hyponym	Synonym	Antonym	Meronym
0	0	*pnꜥ*	*ꜥḥꜥ*	0

Figure 14. Linguistic dictionary for the Coffin Texts

Each word has four fields of identification (representation in hieroglyphs, alphabetical order, transliteration, and translation), two secondary (bibliography and remarks), and four main groups of fields, some of which have already appeared in the database discussed above:

(1) *Grammatical morphology*: fields indicating part of speech.

(2) *Grammatical semantics*: fields with information on semantic features affecting grammar (mainly valency and Aktionsart issues).

[29] Gracia Zamacona 2010b. For the concepts, see Lakoff & Johnson 1980: 30.

[30] http://www.egypto.ulg.ac.be/Ramses.htm.

[31] Hanks 2005.

(3) *Lexical morphology*: fields with information on how words are built in Egyptian, an extremely difficult subject but at the same time crucial for the understanding of the language.

(4) *Lexical semantics*: fields with information on the very abstract subject of word meaning structure. This part will surely require reflection on methodological and theoretical aspects (for example, prototype-based approaches).[32]

BIBLIOGRAPHY

ALLEN, James P. 2006. *The Egyptian Coffin Texts VIII: Middle Kingdom copies of Pyramid Texts*, Chicago (= OIP 112).

ASSMANN, Jan. 1990. Egyptian mortuary liturgies, in: Sarah ISRAELIT-GROLL (ed.), *Studies in Egyptology presented to Miriam Lichtheim*, Jerusalem, I, p. 1-45.

BIBER, Douglas, Susan CONRAD & Randi REPPEN. 1998. *Corpus linguistics. Investigating language structure and use*, Cambridge.

BOHAS, Georges. 1997. *Matrices, étymons, racines. Eléments d'une théorie lexicologique du vocabulaire arabe*, Louvain (= Orbis Supplementa 8).

CT = Adriaan DE BUCK. 1935-1961. *The Egyptian Coffin Texts I-VII*, Chicago (= OIP 24, 49, 64, 67, 73, 81 & 87).

DEPUYDT, Leo. 1985a. À propos de la notion de mouvement en copte et en égyptien, in: *Chronique d'Égypte* LX/ 119-120, p. 85-95.

—. 1985b. On the notion of movement in Egypto-Coptic and Biblical Hebrew, in: Sarah ISRAELIT-GROLL (ed.), *Pharaonic Egypt. The Bible and Christianity*, Jerusalem, p. 30-37.

FRONZAROLI, Pelio. 1973. Statistical methods in the study of Ancient Near Eastern languages, in: *Orientalia* 42, p. 97-113.

GRACIA ZAMACONA, Carlos. 2008. *Les verbes de mouvement dans les Textes des Sarcophages : étude sémantique*. Unpublished PhD dissertation. Paris.

—. 2010a. Space, time and abstract relations in the Coffin Texts, in: *Zeitschrift für Ägyptische Sprache und Altertumskunde* 137, p. 13-26.

—. 2010b. The spatial adjunct in Middle Egyptian: data from the Coffin Texts, in: Kristina LAHN & Maren-Grisch SCHRÖTER (eds.), *Raumdimensionen im Altertum - Zum spatial turn in den Kulturwissenschaften*, *MOSAIKjournal* 1 (Cf. abstract http://www.mosaikjournal.com/).

GRIESHAMMER, Reinhard. 1974. *Die altägyptischen Sargtexte in der Forschung seit 1936: Bibliographie zu de Bucks, The Egyptian Coffin Texts I-VII*, Wiesbaden (= Ägyptologische Abhandlungen 28).

GRUNERT, Stefan & Ingelore HAFEMANN (eds.). 1999. *Textcorpus und Wörterbuch. Aspekte zur ägyptischen Lexikographie*, Leiden (= Probleme der Ägyptologie 14).

GUNDLACH, Rolf & Wolfgang SCHENKEL. 1970. *Lexikalisch-grammatische Liste zu Spruch 335A der altägyptische Sargtexte LL/CT.335A. I-II*, Darmstadt (= Schriftenreihe des Deutschen Rechenzentrums S - 8/1-2).

HAFEMANN, Ingelore. 2001. Zum Zusammenspiel von Semantik und Syntax ägyptischer Verben, in: *Lingua Aegyptia* 10, p. 151-210.

—. 2006. Lexikon und Argumentstruktur, in: *Lingua Aegyptia* 14, p. 349-374.

HANKS, Patrick W. 2005. Typicality and meaning potentials, in: Geoffrey SAMPSON & Diana MCCARTHY (eds.), p. 58-66.

HINTZE, Fritz. 1974. Statistisches zu den Sargtexten, in: *Göttinger Miszellen* 9, p. 63-74.

JUNGE, Friedrich & Wolfgang SCHENKEL. 1972. Göttinger Konkordanz zu den altägyptischen Sargtexten, in: *Göttinger Miszellen* 3, p. 37-38.

LAKOFF, George & Mark JOHNSON. 1980. *Metaphors We Live By*, Chicago.

LÜDELING, Anke & Merja KYTÖ (eds.). 2008. *Corpus linguistics*, I. Berlin – New York.

[32] See a useful introduction to these issues applied to Egyptology in Nyord 2010.

MATHIEU, Bernard. 2004. La distinction entre Textes des Pyramides et Textes des Sarcophages est-elle légitime ?, in: Susanne BICKEL & Bernard MATHIEU (eds.), *D'un monde à l'autre. Textes des Pyramides et Textes des Sarcophages*, Cairo (= Bibliothèque d'Étude 139), p. 247-262.

VAN DER MOLEN, Rami. 2005. *An analytical concordance of the verb, the negation and the syntax in Egyptian Coffin Texts I-II*, Leiden (Handbook of Oriental Studies, Section 1: The Near and Middle East 77).

NYORD, Rune. 2010. Radial structure of Middle Egyptian prepositions, in: *Zeitschrift für Ägyptische Sprache und Altertumskunde* 137, p. 27-44.

PEUST, Carsten. 2007. Die Konjugation des Verbs für gehen im Neuägyptischen, in: *Göttinger Miszellen* 212, p. 67-80.

VAN DER PLAS, Dirk & Joris F. BORGHOUTS. 1998. *Coffin Texts Word Index*, Utrecht/Paris (= Publications Interuniversitaires de Recherches Égyptologiques Informatisées VI).

RISSANEN, Matti. 2005. On the history of that/zero as object clause links in English, in: Geoffrey SAMPSON & Diana MCCARTHY (eds.), p. 137-148.

SAMPSON, Geoffrey & Diana MCCARTHY (eds.). 2005. *Corpus linguistics. Readings in a widening discipline*. London/New York.

SCHENKEL, Wolfgang. 1982. Eine Konkordanz zu den altägyptischen Sargtexten, in: Jean LECLANT (ed.), *L'Égyptologie en 1979*, Paris, II, p. 45-53.

—. 1983. *Aus der Arbeit an einer Konkordanz zu den altägyptischen Sargtexte. Teil I: Zur Transkription des Hieroglyphsch-Ägyptischen. Teil II: Zur Pluralbildung des Ägyptischen*, Wiesbaden (= Göttinger Orientforschungen IV/12).

—. 1988. Sprachforschung und Textquellen. Integrierte Datenverarbeitung als konkrete Utopie, in: Sylvia SCHOSKE (ed.), *Akten des vierten internationalen Ägyptologen Kongresses, München 1985*, Hamburg, III, p. 1-27.

—. 1994. Das Tübinger Konkordanz-Programm, in: *Zeitschrift für Ägyptische Sprache und Altertumskunde* 121, p. 142-153.

SCHENKEL, Wolfgang & W.F. REINEKE. 1998. Textcorpus und Wörterbuch. Arbeitstagung aus Anlass des Neubeginns der Arbeit am ägyptischen Wortschatz ein Jahrhundert nach der Gründung der akademischen Kommission zur Herausgabe des 'Wörterbuchs der ägyptischen Sprache', Berlin, 22.-26. September 1997, in: *Göttinger Miszellen* 162, p. 107-109.

VERNUS, Pascal. 1990. *Future at Issue: Tense, Mood and Aspect in Middle Egyptian: Studies in Syntax and Semantics*, New Haven (= Yale Egyptological Studies 4).

—. 1996. La position linguistique des Textes des Sarcophages, in: Harco WILLEMS (ed.), *The World of the Coffin Texts*, Leiden (= Egyptologische Uitgaven 9), p. 143-196.

WENTE, Edward F. 1959. *The Syntax of Verbs of Motion in Egyptian*, Unpublished PhD dissertation, Chicago.

WILLEMS, Harco. 1988. *Chests of Life. A Study of the Typology and Conceptual Development of Middle Kingdom Standard Class Coffins*, Leiden.

WINAND, Jean. 1987. *Le voyage d'Ounamon. Index verborum. Concordance. Relevés grammaticaux*, Liège (= Ægyptiaca Leodinensia 1).

—. 1990. Les bases de données de textes en égyptien, in: *Informatique et Égyptologie* 7, p. 161-169.

—. 1991. Le verbe *ỉy* / *ỉw* : unité morphologique et sémantique, in: *Lingua Aegyptia* 1, p. 357-387.

—. 2006. *Temps et aspect en égyptien : une approche sémantique*, Leiden (= Probleme der Ägyptologie 25).

The Digital Library of Inscriptions and Calligraphies

Azza EZZAT
Alexandria

1. INTRODUCTION

Digital publications have become one of the means adopted by the world today for preserving cultural and historical heritage. To this end, the Bibliotheca Alexandrina Calligraphy Center has established an electronic project which documents and publishes different ancient inscriptions. This project is called 'The Digital Library of Inscriptions and Calligraphies", and its development is the primary goal of the Calligraphy Center, which has taken upon itself the publication of different calligraphy styles and inscriptions, especially those left behind by the various languages that influenced Egypt. The center makes all of this available to scientists, scholars, and the public in the form of simple digital content on the internet.

2. THE PROJECT OF THE DIGITAL LIBRARY

The Digital Library of Inscriptions and Calligraphies[1] (DLIC) is a digital archive for all types of inscriptions on all kinds of supports from all historical periods. These inscriptions are displayed on the website of the DLIC in digital form, which includes images and a brief description for each. The DLIC launched 1500 inscriptions on 13 August 2010 and, less than one year later, this number now exceeds 4000 inscriptions, all available online for free.

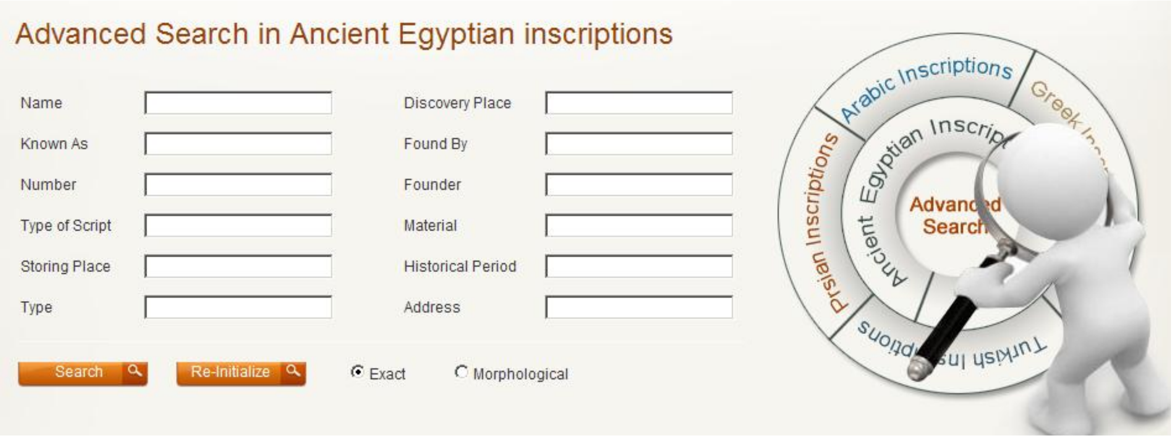
The website of the Digital Library of Inscriptions and Calligraphies (see n. 1) has been developed in collaboration with the International School of Information Science[2] (ISIS); a research institute affiliated with the Bibliotheca Alexandrina which initiates, promotes and incubates IT projects and activities related to building a universal digital library.

The website is designed to provide easy access of researchers to the valuable collection, making available the images and references for each inscription. Users can browse easily through inscriptions on the DLIC, as they are categorized according to the original language of the inscription, the classification and type of support.

The advanced search facility enables users to search using the registration number of the inscription, the place where it is kept, the place where it was found, or the historical period to which it belongs (see Fig. 1). As such, researchers can retrieve the full data related to the inscription: high-resolution images, analyses of the inscription, various types of metadata and a brief description of the object, in addition to a translation of the inscription.

[1] http://inscriptionslibrary.bibalex.org. The official opening of the website was on the 16th of October 2012 (10th anniversary of the inauguration of the Bibliotheca Alexandrina). For further information, contact DLIC@bibalex.org.

[2] http://www.bibalex.org/isis.



Advanced Search in Ancient Egyptian inscriptions

Name		Discovery Place	
Known As		Found By	
Number		Founder	
Type of Script		Material	
Storing Place		Historical Period	
Type		Address	

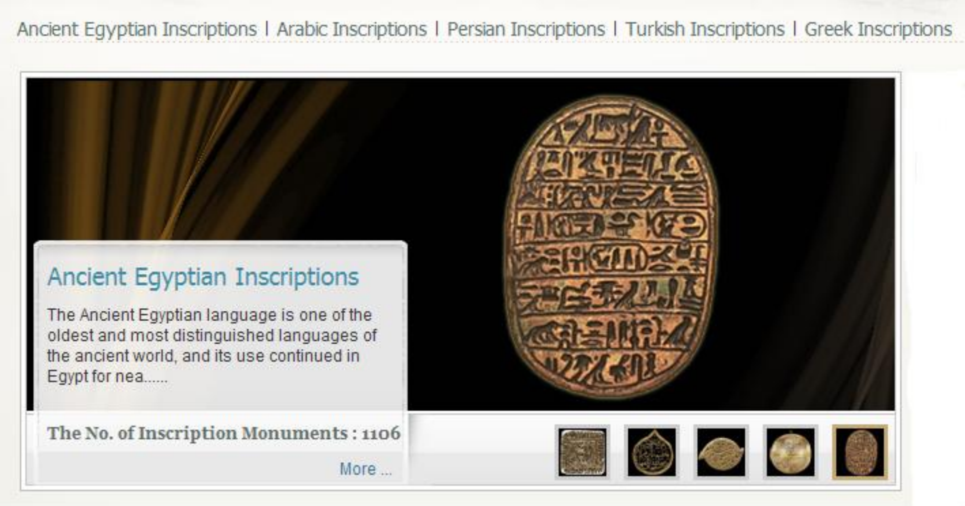
Search Re-Initialize ● Exact ○ Morphological

Figure 1. Advanced search in the DLIC

As for the Technical details, ISIS has already designed and implemented a database for metadata, transliterations, translations and pictures. A workflow was implemented and is being used for documenting monuments' metadata and their corresponding inscriptions. To date, 5000 objects, associated with their metadata and inscriptions, have been documented. A web-based application with both Arabic and English interfaces was implemented and tested. The team has designed a web interface that has been published in October 2012 (see n. 1).

3. The content of the Digital Library of Inscriptions and Calligraphies

In its initial stage, the project began with the documentation of the calligraphies of the group of languages that comprises Ancient Egyptian, Arabic, Persian, Turkish, and Greek languages. This is complemented by another group of various calligraphies, namely, the Thamudic, the Nabataean, and the Musnad. As the documentation of the calligraphy of each group of languages is completed, work begins on a new group of languages.



Ancient Egyptian Inscriptions | Arabic Inscriptions | Persian Inscriptions | Turkish Inscriptions | Greek Inscriptions

Ancient Egyptian Inscriptions

The Ancient Egyptian language is one of the oldest and most distinguished languages of the ancient world, and its use continued in Egypt for nea......

The No. of Inscription Monuments : 1106

More ...

Figure 2. Navigating the calligraphies of the included languages

3.1. *The Ancient Egyptian Language*

The language of the ancient Egyptians passed through several phases, each of which had a different form of writing. It started with the Hieroglyphic and Hieratic scripts. After that came Demotic, a shorthand type of writing used for the purposes of daily life. Finally, the Coptic script appeared; it is the result of a long contact with the Hellenic civilization and is a mixture of Greek and Demotic letters. Each of these four scripts, which merged under the umbrella of the Ancient Egyptian language, is attested by many monuments and artifacts that reflect the development of the ancient Egyptian language.

3.2. *The content of the Ancient Egyptian language section in the DLIC*

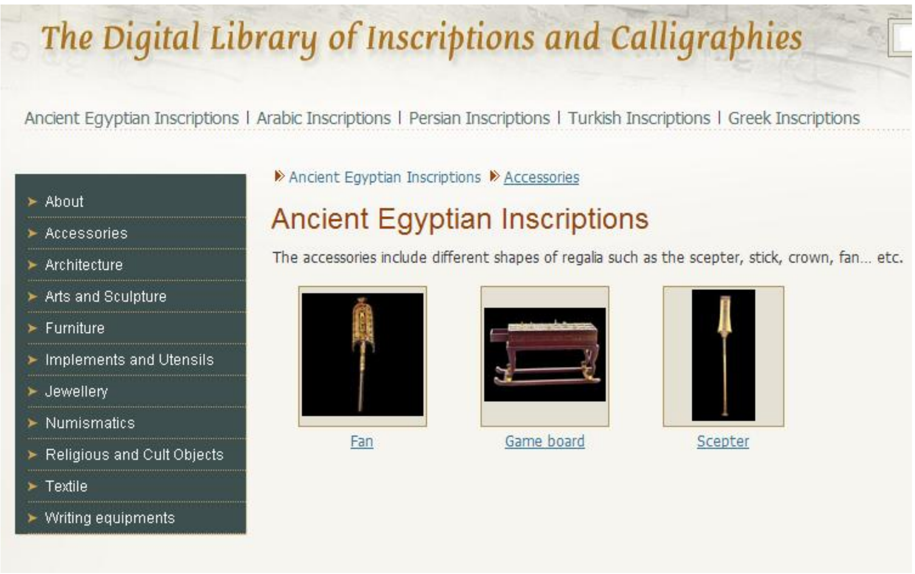
The Ancient Egyptian language section is classified into ten main categories, each of them comprising various types of objects.

3.2.1. Accessories

This category in the Ancient Egyptian languages includes different shapes of regalia such as scepters, sticks, crowns, fans, etc.

Ex. 1 The Senet game-board of king Tutankhamun.

Ex. 2 The funerary fans and royal sceptres which were found in the tomb of king Tutankhamun.



The Digital Library of Inscriptions and Calligraphies

Ancient Egyptian Inscriptions

The accessories include different shapes of regalia such as the scepter, stick, crown, fan... etc.

Ancient Egyptian Inscriptions | Arabic Inscriptions | Persian Inscriptions | Turkish Inscriptions | Greek Inscriptions
About
Accessories
Architecture
Arts and Sculpture
Furniture
Implements and Utensils
Jewellery
Numismatics
Religious and Cult Objects
Textile
Writing equipments
Ancient Egyptian Inscriptions
Accessories
Fan
Game board
Scepter

Figure 3. Ancient Egyptian inscriptions section in the DLIC

3.2.2. Architecture

The digital library also includes a set of written inscriptions recorded on various types of Ancient Egyptian architectural elements. Amongst them are a set of:

Ex. 1 Obelisks, such as Hatshepsut's obelisk at Karnak temple.

Ex. 2 Wall inscriptions, such as the pyramid texts engraved in the Unas pyramid and the Kadesh battle reliefs recorded at the Ramesseum.

Ex. 3 Tombs, such as that of Queen Nefertari.

Ex. 4 False doors, such as those of Ptahshepses.

Ex. 5 Lintels, such as those of Ramesses II from a temple at Abydos.

Ex. 6 Royal stelae, such as that of Merenptah ("Israel stela") or the dream stela of Tuthmosis IV.

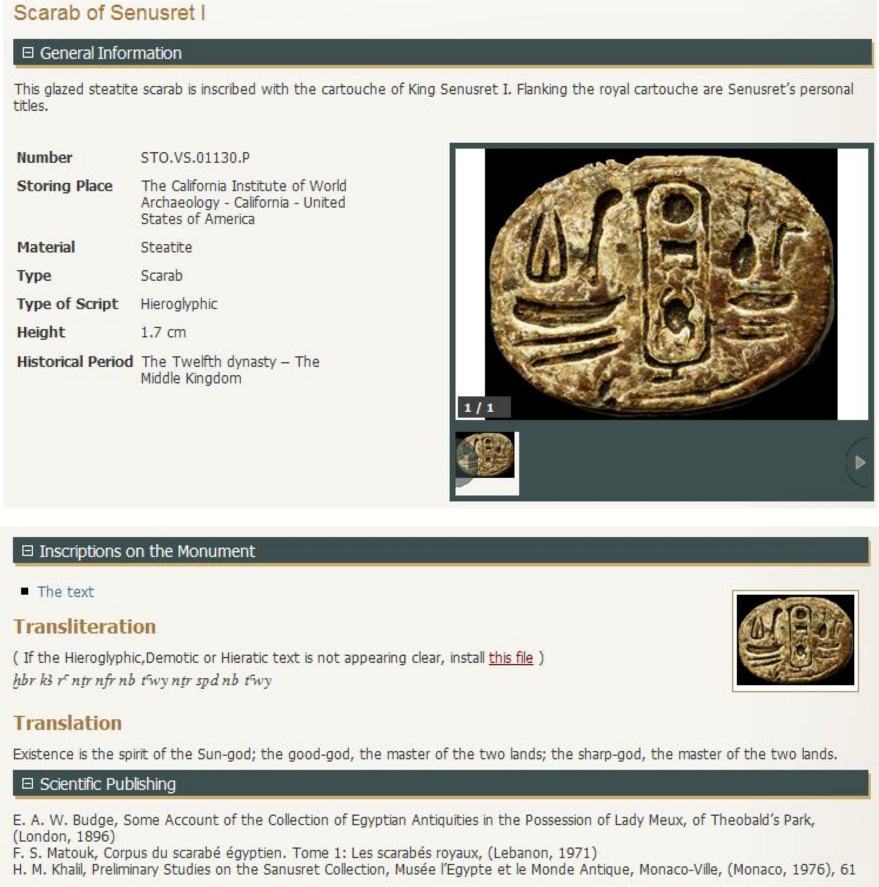
Ex. 7 Stelae of individuals, such as that of Nit-ptah.

3.2.3. Arts and sculpture

The arts and sculpture category covers a cluster of statues in different shapes and sizes:

Ex. 1 The group statue of Menkaure, Ra-hotep and Nofert.

Ex. 2 Seated statue of Djoser.
Ex. 3 Statuette of King Khufu.
Ex. 4 Sphinx statues, such as those of Hatshepsut.
Ex. 5 Scarabs, such as those belonging to king Sensusert I, Tuthmosis III and Amenhotep III.



Scarab of Senusret I

⊟ General Information

This glazed steatite scarab is inscribed with the cartouche of King Senusret I. Flanking the royal cartouche are Senusret's personal titles.

Number	STO.VS.01130.P
Storing Place	The California Institute of World Archaeology - California - United States of America
Material	Steatite
Type	Scarab
Type of Script	Hieroglyphic
Height	1.7 cm
Historical Period	The Twelfth dynasty – The Middle Kingdom

⊟ Inscriptions on the Monument

- The text

Transliteration

(If the Hieroglyphic,Demotic or Hieratic text is not appearing clear, install this file)
ḫbr kȝ rʿ nṯr nfr nb tʿwy nṯr spd nb tʿwy

Translation

Existence is the spirit of the Sun-god; the good-god, the master of the two lands; the sharp-god, the master of the two lands.

⊟ Scientific Publishing

E. A. W. Budge, Some Account of the Collection of Egyptian Antiquities in the Possession of Lady Meux, of Theobald's Park, (London, 1896)
F. S. Matouk, Corpus du scarabé égyptien. Tome 1: Les scarabés royaux, (Lebanon, 1971)
H. M. Khalil, Preliminary Studies on the Sanusret Collection, Musée l'Egypte et le Monde Antique, Monaco-Ville, (Monaco, 1976), 61

Figure 4. Scarab of Senusret I: main data & inscription on the object

3.2.4. Furniture

Chairs, headrests, and boxes are the most important elements of the furniture category of the digital library project.

Ex. 1 The gilded chair of princess Sitamen.
Ex. 2 The funerary headrests of Tutankhamun made of glass and ivory.
Ex. 3 The boxes of king Tutankhamun.

3.2.5. Implements and utensils

The implements and utensils section include different types of equipment, such as those used for music, warfare, hunting, agriculture, and cosmetics, in addition to some games and some equipment used for measuring distance, time, and volume:

Ex. 1 Musical equipment, such as the clappers of Sit-Hathor.
Ex. 2 Hunting equipment, such as that of Tutankhamun.
Ex. 3 Cosmetic equipment, such as kohl pots.
Ex. 4 Vessels, such as the vases belonging to king Amenhotep II and Tutankhamun, as well as the pilgrimage flask of Saint Mena.

3.2.6. Jewelry

Distinguished by its diversity and richness, the collection of Ancient Egyptian inscriptions includes a diverse collection of jewelry comprising various types pectorals, necklaces, rings and bracelets.

Ex. 1 Jewels of king Tutankhamun.
Ex. 2 Jewels of king Tuthmosis IV.
Ex. 3 Tanis treasure which belonged to king Psusennes I and Sheshonq II.

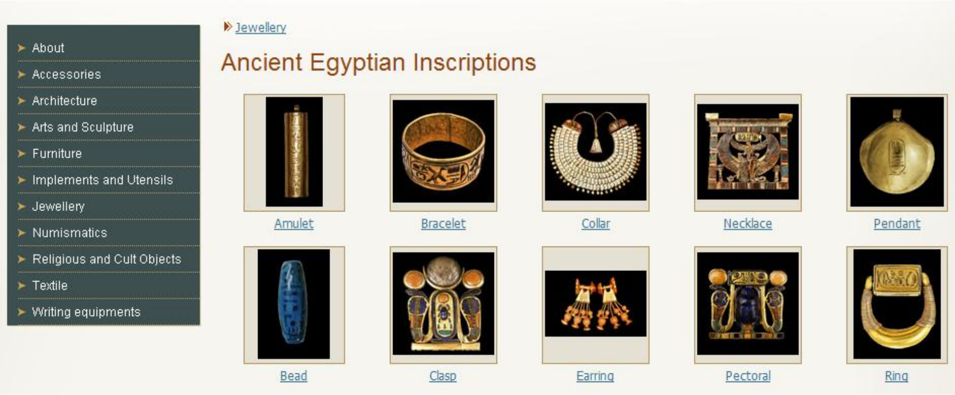


Figure 5. The jewelry section in the DLIC

3.2.7. Religious and cult objects

The religious and cult objects in the Digital Library comprise more than 485 objects such as tombstones, crosses, cones, and mummy labels.

Ex. 1 Ushabti statuettes, such as those of Senndjem and king Psmatik I.
Ex. 2 Canopic jars and chests, such as those of Tutankhamun and Hatshepsut.
Ex. 3 Coffins, such as the coffins of the two brothers Khnum-nakht and Nakht-ankh.
Ex. 4 Unique Coptic icons and tombstones preserved at the Coptic Museum in Cairo.

3.2.8. Textile

The Ancient Egyptian textile section contains a set of funerary textiles.

Ex. 1 The textile which was found in the tomb of Tuthmosis IV.
Ex. 2 Tutankhamun's wardrobe and linen wrapping.
Ex. 3 Coptic tapestry.

3.2.9. Coinage

Coinage in the Digital Library project contains a copy of the oldest coin attested from Ancient Egypt, dating back to the reign of king Nektanebo II and preserved in the British Museum.

3.2.10. Writing equipment

Additionally, the digital library has a set of writing equipment that comprises around 250 inscriptions recorded on a set of:

Ex. 1 Scribal palettes.
Ex. 2 Ostraca inscribed in Hieratic, Demotic, and Coptic scripts.
Ex. 3 Papyri with copies of texts such as *The Shipwrecked Sailor*, *Sinuhe*, and *The Eloquent Peasant*.

The *AGÉA* Database Project

Anthroponymes et Généalogies de l'Égypte Ancienne

Yannis Gourdon
Institut Français d'Archéologie Orientale

1. Introduction

Since the 30s, our understanding of ancient Egyptian personal names has been dependent on a very valuable resource: Ranke's *Personennamen*. This dictionary includes about 14 200 names, which are mostly analysed, translated and documented. However, 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 in the field.

This was one of the aims of the author's Ph.D. dissertation on Egyptian personal names, which focused on the Old Kingdom,[1] a period which until now has not been dealt with in a global study.[2] But this piece of work was not only a new version of the *PN* for the Old Kingdom, it was also intended to produce a thorough study of the grammatical structures and the sociological phenomena related to the personal names, like the transmission of names within families, their political and religious content, and so on.

It was important to publish it both within a reasonable time-frame and in an appropriate format that would at the same time be accessible to all and easy to update. For these reasons, I chose to publish the dictionary as an online database. The *AGÉA* database project[3] presented here directly proceeds from this idea.

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. This database is to be regarded as a tool facilitating more efficient analysis and a better interpretation of data.

To gain a better idea of the social practices connected with Egyptian personal names, the genealogical and prosopographical data is indispensable, which explains their inclusion in *AGÉA*. Nevertheless, this database is not to be regarded as strictly prosopographical, in particular as far as the inclusion of titles in transliteration only is just a tool intended to clear up onomastic issues. For all these reasons, the name *AGÉA* (*Anthroponymes et généalogies de l'Égypte ancienne*) was given to this

[1] *Recherches sur l'anthroponymie dans l'Égypte du III^e millénaire avant J.-C. : signification et portée sociale du nom égyptien avant le Moyen Empire* (defended on the 5th of January 2007, at the University Lumière-Lyon 2, under the supervision of Prof. L. Pantalacci).

[2] The forthcoming publication of Katrin Scheele-Schweitzer's *Die Personennamen des Alten Reiches: altägyptische Onomastik unter Lexikographischen und Sozio-kulturellen Aspekten (PN-AR)* also goes in this direction.

[3] It is included in the anthroponomastic part of the onomastic research program I lead with Å. Engsheden at the French Institute in Cairo.

database, though it is very different from former databases like Jochen Hallof's *Prosopographia Aegypti* at the CCER.

Connected with my own researches, the first development of *AGÉA* concerns the Old Kingdom. Every single name and person will be recorded as far as they are known, including all kinds of published or unpublished sources (funerary material, rock inscriptions, hieratic archives, etc.).

2. Data and the FilemakerPro design of AGÉA

2.1. *Data and sections*

All the data are first collated in a FileMakerPro database before being transferred to a PostgreSQL software program designed for web access. In *AGÉA*, data are distributed over two sections: names and individuals.

2.1.1. The names section

The names section is divided into two parts. The first part contains basic data about names, showing every attested writings of a name and every person who bears it. The second part contains data concerning grammatical analysis of each name.

First part (basic data)

- **PN number**. If the name is registered in H. Ranke's dictionary, his PN number is quoted.
- **Hieroglyphic name writings.** The hieroglyphic transcription is made with S. Rosmorduc's Jsesh software, using MacScribe font. It also includes a hieroglyphic encoding field based on the 'Manuel de Codage'. For the sake of convenience and because the purpose of this database is not strictly palaeographical, the writings of names are systematically given in a hieroglyphic transcription, even when names were originally written in hieratic. One always attempts to give all graphic variants of any name, even if these belonged to the same person. However, showing all written forms is not considered a necessity, since not all writings are variants. Hence, it was essential to define what was a unique written form. Obviously, reversion of the orientation of a single sign or the placement of one sign over another does not constitute a significant variation. On the other hand, inversion between two or several signs, the use of particular determinatives, haplography, various signs for the same phonetic value, or any curious written form can be interesting for the reading of names. The addition of a semi-consonant to a *Vollname*[4] can also hold some interest, but it is advisable to be sceptical of some phonetic complements that are not always significant, for example in the case of a *yod* used as a determinative.[5]
- **Transliteration of the name.** The transliteration is a simplified one, using dots to mark suffixes and grammatical forms, but neither the feminine nor the plural. Each word composing the personal name is separated from others words by hyphens. This standard process is necessary to individualize the names within a transliterated text.
- **Translation of the name.** The translation of names is mostly a personal one based on the author's study of Old Kingdom names. This translation between quotation marks doesn't show hyphens, because in a translation of a text the name is not to be translated and thus does not need to be distinguished from the rest of the text.
- **Sex of the name bearer.** This is indicated by "h." for "homme", "f." for "femme", as the database is originally in French. These two letters are followed by a question mark if a doubt

[4] Vernus (1989: 145-161). For the expression "Vollname", see PN II: 20-94.

[5] Clère (1959: 76-78).

remains. The question mark is used alone when the identity of the carrier of the name cannot be established.

- **Bibliography.** Bibliographical references are given for each hieroglyphic writing of a name related to a single individual.
- **Dating of the name bearer**. In this kind of database, there is not room for discussion about every piece of data used to date each individual. However, it should be noted that all common dating criteria for the Old Kingdom were used: Cherpion's dating criteria,[6] royal name, archaeological data, etc.
- **"Origin" of the name bearer.** This field sums up all data related to the origin of the objects mentioning a single person.
- **Comment.** This is a general comment about the name, usually connected to its attestation in the *PN*.

Second part (grammatical data)[7]

The second part of the names section includes a complete description of each name, from a linguistic and onomastic point of view. It also contains a grammatical comment field about the name, mostly dealing with the different readings proposed by other scholars.

2.1.2. The individuals section

The individuals section contains all the data about each person registered in the database. Every individual receives a *personal numbered name* in transcription (as 'Shepsespouptah 1') in order to distinguish homonymous names; the ID is random.

- **Each name** borne by one person is mentioned with its specification (*rn.f nfr*, *rn.f ꜥꜣ*…).
- **Dating** and **origin** of the individual.
- **Titles** are given in transliteration only, with their corresponding index number, when there is one (e.g. "Jones, 1382"[8] for the title *wꜥb nswt* in the Old Kingdom).
- **All parents** are mentioned with the indication of their relationship with the individual. If the relationship formula (*zꜣ.f smsw*, *mwt.s*, *snt.f*, etc.) is known in the original documentation, it is quoted.
- **Genealogical trees** can be provided for the largest families.
- **Comment.** This field includes a comment about the individual and his family.

2.2. *The FileMakerPro design and tables*

At the beginning, all data included in the author's PhD corpus were collated in a FileMakerPro 5 to 7 database. When it was decided to create an online database at the IFAO, it was necessary to update these previous files and restructure them to facilitate transfer of data to PostgreSQL software, which is used at the French institute for all online databases. So the database was completely redesigned with a FileMakerPro 8 structure, which is based on a pivot-tables system and allows the creation of several-to-several links and a tool for performing cross-references.

This FileMakerPro database uses 4 (or 5, where "other names" are applicable) major tables, which are:

- **hieroglyphic names writings**, connected with the "names" table by a pivot-table;
- **names**, linked to the "individuals" table by a pivot-table;

[6] Cherpion (1989). See also Baud (1998: 31-95).

[7] This part of the names section will be available only in the second phase of the launch of *AGÉA* on the internet.

[8] Jones (2000: n° 1382).

- **individuals**, directly connected with the "**other names**" table (for other names borne by a single individual);
- and **a cross-references table**, which is a pivot-table between these three/four previous tables and the bibliographical table mentioned below.

Beside these major tables are others, which include, for instance, secondary data for the study of names:

- **titles**, linked to the "individuals" table by a pivot-table;
- **parents**, which are actually defined through an internal link within the "individuals" table and via a pivot-table;
- and **bibliography**.

In all these tables, each key item, such as names, names writings, individuals, titles, receives a set ID-number. This number is given simply to identify these items with no connection with anything else at all. The main table used to fill the database is the cross-reference table, linked to the essential bibliographical/writings and individual tables. From this table it is possible to fill the most significant data within other tables.

3. WEB ACCESS TO AGÉA

A beta version of the *AGÉA* database is already available for free public access on the IFAO website (www.ifao.egnet.net/bases/agea/noms/).

For the web-accessible version, *AGÉA* follows a common format used for all IFAO databases: it is an open-source PostgreSQL database using a PHP language.[9]

As *AGÉA* uses a number of conventions for the transliteration, transcription and presentation of the names, an introductory page presents these conventions with a short history of anthroponomastic studies in Egyptology.

3.1. *The PostgreSQL design*

Like all the IFAO web-accessible databases, *AGÉA* follows a common presentation, thus *AGÉA* will open on the list of names.[10] This list of names presents:

(1) The *name-ID*.
(2) The more complete writing known (as an example).
(3) Transliteration.
(4) A French translation of the name.
(5) The *PN* number.

Each name-ID and name writing has a hypertext link, which opens the related name file. Its content is the following:

(1) The 5 previous data (repeated).
(2) The translation of the name, where possible.
(3) Each reference is introduced by a name writing number with the connected name writing and its hieroglyphic encoding.
(4) The personal numbered name.

[9] I am very grateful to Chr. Gaubert of the IFAO IT service, who designed the web-accessed version of *AGÉA*. He always tried to stick to my recommendations with the possibilities offered by the PostgreSQL software.

[10] I am presently developing an English version of *AGÉA*.

(5) The sex of the name bearer.
(6) The origin and dating of his monuments.
(7) The relevant bibliographical references for the writing mentioned above.
(8) A general comment field about the name and its attestation in the *PN*.

The *personal numbered name* itself has a hypertext link that opens the file of this person. Its content is the following:

(1) The individual-ID, followed by some of the previous basic data (4, 5, and 6).
(2) Titles he bears with their index numbers (Jones' for the Old Kingdom).
(3) The list of all his relatives with indications of their relationship to the individual. Each relationship link written in the original documentation is quoted.[11]
(4) A comment field about the individual and his family.

For each relative (individual-ID or *personal numbered name*) a hypertext link gives access to their file.

Like names, all individuals are listed. This page sums up the essential data about each individual (ID, *personal numbered name*, sex, origin and dating).

3.2. *The queries on PostgreSQL*

The primary function of any database is to make queries. In the names section of *AGÉA*, one can search for a name in transliteration, transcription, by its *PN* number or its *AGÉA* ID. One can also search by dating. In this case, a list of all the dates (attested on monuments) filed in the database appears as a drop-down menu. The same principle is applied for the origin query. All these kinds of queries can be done for the individuals section of *AGÉA*. But in this section, one can also search by titles or index number.

Finally, one of the greatest benefits of the presence of a 'Manuel de Codage' field in the *AGÉA* database is to enable searches to be conducted within strings of hieroglyphs. For example, if one wants to know which names contain the A4 sign of Gardiner's sign-list, one just has to write A4 in the search field and the relevant names will be displayed.

4. CONCLUSION

At this point in time, the FileMakerPro version of the *AGÉA* database includes about 4 200 names and 11 500 persons. These data need be treated so that they can be included in the online version, but I hope that most of the entire Old Kingdom corpus will be registered soon. A first version of the *AGÉA* database is now available on the IFAO website.

Collaborations with some French teams (the MafS at South-Saqqara, the IFAO at Tabbet el-Guech and Balat) already led to the incorporation of many unpublished data. Contacts were made with Australian, Czech and Polish colleagues and I hope I expand these contacts and collaborations to others teams, including Egyptian colleagues who dig in Abusir, Saqqara and Giza.

As mentioned in the introduction, the *AGÉA* database project will not end with the Old Kingdom. Contacts and collaborations have already been made to extend *AGÉA* to other pharaonic periods. For example, I am working with M. Borla (Egyptian Museum of Turin) to include data from her Imy-renef database[12] in *AGÉA*. I am also working with M. Thirion to include her very valuable corpus of theophoric names[13] in *AGÉA*. Further, a collaboration with N. Favry (Paris IV-Sorbonne University) and

[11] For the largest families, a genealogical tree will be provided.

[12] From her unpublished PhD. dissertation *Répertoire onomastique et prosopographique des fonctionnaires thébains du Nouvel Empire et de la Troisième période Intermédiaire d'après les documents conservés au Musée Égyptien de Turin (oushebtis et objets en relations).*

[13] See her articles in *Revue d'Égyptologie* quoted in the bibliography.

with M. Marée (British museum) is being set up for titles and genealogical data of the Middle Kingdom.

BIBLIOGRAPHY

BAUD, Michel. 1998. À propos des critères iconographiques établis par Nadine Cherpion, in: Nicolas GRIMAL (ed.), *Les critères de datation stylistiques à l'Ancien Empire*, Cairo, IFAO, p. 31-95 (= *Bibliothèque d'Étude* 120).

CHERPION, Nadine. 1989. *Mastabas et hypogées d'Ancien Empire. Le problème de la datation*, Bruxelles.

CLÈRE, Jacques Jean. 1959. L'emploi du signe du roseau (*ỉ*) comme déterminatif dans l'écriture égyptienne, in: Frank HERBERT (ed.), *Akten des XXIV Internationalen Orientalisten-Kongress, München 28 August bis 4 September 1957*, Wiesbaden, p. 76-78.

JONES, Dilwyn. 2000. *An Index of Ancient Egyptian Titles, Epithets and Phrases of the Old Kingdom*, Oxford, 2 vols.

PN = Hermann RANKE. 1935; 1949-1952; 1977†. *Die ägyptischen Personennamen*, Glückstadt, 3 vols.

THIRION, Michèle. 1979. Notes d'onomastique, contribution à une révision du Ranke *PN*, in: *Revue d'Égyptologie* 31, p. 81-96.

—. 1981. Notes d'onomastique, contribution à une révision du Ranke *PN* (deuxième série), in: *Revue d'Égyptologie* 33, p. 79-87.

—. 1982-1983. Notes d'onomastique, contribution à une révision du Ranke *PN* (troisième série), in: *Revue d'Égyptologie* 34, p. 101-114.

—. 1985. Notes d'onomastique, contribution à une révision du Ranke *PN* (quatrième série)", in: *Revue d'Égyptologie* 36, p. 125-143.

—. 1986. Notes d'onomastique, contribution à une révision du Ranke *PN* (cinquième série), in: *Revue d'Égyptologie* 37, p. 131-137.

—. 1988. Notes d'onomastique, contribution à une révision du Ranke *PN* (sixième série), in: *Revue d'Égyptologie* 39, p. 131-146.

—. 1991. Notes d'onomastique, contribution à une révision du Ranke *PN* (septième série), in: *Revue d'Égyptologie* 42, p. 223-240.

—. 1992. Notes d'onomastique, contribution à une révision du Ranke *PN* (huitième série), in: *Revue d'Égyptologie* 43, p. 163-168.

—. 1994. Notes d'onomastique, contribution à une révision du Ranke *PN* (neuvième série), in: *Revue d'Égyptologie* 45, p. 175-188.

—. 1995. Notes d'onomastique, contribution à une révision du Ranke *PN* (dixième série), in: *Revue d'Égyptologie* 46, p. 171-186.

—. 2001. Notes d'onomastique, contribution à une révision du Ranke *PN* [Onzième série], in: *Revue d'Égyptologie* 52, p. 265-276.

—. 2003. Notes d'onomastique, contribution à une révision du Ranke *PN* [Douzième série], in: *Revue d'Égyptologie* 54, p. 177-190.

—. 2004. Notes d'onomastique, contribution à une révision du Ranke *PN* [Treizième série], in: *Revue d'Égyptologie* 55, p. 149-159.

—. 2005. Notes d'onomastique, contribution à une révision du Ranke *PN* [Quatorzième série], in: *Revue d'Égyptologie* 56, p. 177-190.

VERNUS, Pascal. 1989. La stèle du pharaon *Mnṯw-ḥtpi* à Karnak : un nouveau témoignage sur la situation politique et militaire au début de la D.P.I., in: *Revue d'Égyptologie* 40, p. 145-161.

Computers and Journal Publishing[*]

A Position Paper

Eugene Cruz-Uribe

Editor of the *Journal of the American Research Center in Egypt*

The issue presented in this position paper relates to the ever changing nature of journal publication in the world of electronic media and how it affects print media. Many issues need to be addressed and technology is often driving changes whether users wish them or not. In the following I present some comments that I hope will engender further discussion and hopefully will assist us in the development of consistent approaches to an ever changing system. The intent is to stimulate the discussion of issues related to publishing articles in Egyptology journals in the age of computers.

We all have our own stories of the value and ease of using computer technology versus older technologies. In my own case my dissertation in 1983 was the first submitted to the Oriental Institute at the University of Chicago that was printed on the University's mainframe computer. While that does not seem like a major item, it spelled the death knell of the university's dissertation office where the mafia of dissertation typists disappeared within a year. And why not? For a starving graduate student paying someone $1-2 to have a page of a dissertation typed according to a specific format versus having the mainframe programmed to do it automatically for free was a blessing. But it did warn me that there was a human cost in the use of computer technology.

Nearly 30 years later we are hardly surprised at the advances that have been made in computer software and hardware. To use a computer as a drawing tablet to produce a more accurate drawing of objects and temple scenes (or even the Demotic graffiti I study), place it within a document and from there add all the necessary formatting commands with the simple push of a few buttons may have seemed like magic a few years ago, but is common place now. Or simply taking a document, converting it to a `.pdf` file, and posting it on the internet where it is instantly available worldwide changes the way we can and must approach things.

This brings us directly to the topic of this position paper which is to explore the issues which face editors of Egyptology journals in this technological age. Should we as scholars and editors adopt every technology or should we stand back pick and choose? Should we politely tell our retired colleagues raised in an age of typewriters that no we will not accept that great article because it was produced on an old typewriter and the hieroglyphs have been carefully inserted by hand in ink within the text? What about when our publishers tell us that we must adopt Unicode standards and thus use Unicode hieroglyphs for the preparation of articles they will print? Who really is to have control? Is it the author of the article? Is it the editor of the journal? Or is it the publisher who handles the technical aspects of providing the final printed copy? Or is it a combination of all three?

[*] With the assistance of Julie Masquelier-Loorius (*Revue d'Égyptologie = RdÉ*), Lutz Popko (*Zeitschrift für ägyptische Sprache und Altertumskunde = ZÄS*) and John Gee (*Journal of the Society of the Study of Egyptian Antiquities = JSSEA*).

Another area of interest may be the issue of the World Wide Web. The development of the Internet over the past 20 years has not been an easy phenomenon to watch as a scholar. The Web is by definition "open" with all the wrinkles and problems which that can expose. It has meant every wacko and pyramidiot can and has posted reams of really bad and misleading items for everyone to access. There has been little in the way of filtering material. Yet at the same time, scholars can take material and post them online for everyone to see. This has sometimes lead to the circumstance where certain individuals "publish" their works in the Web media, where it is then picked up by the print and TV media, and then presented as "proven" facts. In the past our discipline has always maintained the necessity of using peer review to carefully evaluate the efficacy and value of proposed changes. How are we to view all those golden mummies from Bahariya when they are never presented to the scholarly community for review, but are printed in coffee table books? What is the significance of those items discovered on Zahi Hawass's behalf for our field, if we have never really evaluated them?

Does this mean that the internet is a bad thing? Does it mean that journals are now going to be replaced by a totally open forum of the world? How will scholars be evaluated? Will it be by peers or will it be by blog or tweets? Where do academic journals fit into that role? Will an online journal automatically be suspect simple because it is online? Or will we as scholars recognize the merits of online publishing and treat it the same as the *Journal of Egyptian Archaeology* (*JEA*) or the *Zeitschrift für ägyptische Sprache und Altertumskunde* (*ZÄS*)?

What about ownership of the printed word (and I use that term very loosely in the context of the World Wide Web)? For those of us who spend significant time in the classroom the expansion of material on the Web has made our jobs correcting student papers extremely difficult. When one can cut and paste so simply, many students despite constant warnings conceive of the Web as not owned by anyone and therefore what is there is actually their own work (whether they had anything to do with producing it or not), simply because they touched a mouse and moved it to their electronic work and provided it to their professor. They made it they say. As professors we say: no you did not, but it does set the stage for us as scholars to try and define within that context what ownership is.

For members of our field we seldom have cases where scholars have plagiarized other's works, but we do have colleagues who wish to post their published articles on the Web. They created them they say. What if the copyright is held by the journal or by the publisher? Does the creator own it or does the "producer"? What if the journal owns the copyright but withholds permission for the author to post it on the Web? Can the author stretch the concept of "fair use" to say he needs to post the entire article on the Web regardless if he/she has permission of the copyright holder? What if the journal has been stored on a service like JSTOR? What effect does that have on ownership issues? As Lutz Popko has noted, the copyright laws themselves do not always provide much help. In Germany, federal law requires copyright to reside with the publisher. In England, it stands with the author. With my journal (the *Journal of the American Research Center in Egypt*) it resides with the journal.

And finally, where are we going in the future? Can we forecast what our field will be like in ten or twenty years? Will we even bother with print journals anymore? Will the idea of a hard copy library have become obsolete? Should we embrace all of the technologies we have now and others that may come in the future? Should we plan for the future or simply let it happen? Or should we sit at academic conferences, where we can meet as friends and discuss the issues *ad nauseam* and then wake up in the morning without having taken any stand on the issues, while circumstances (in this case modern technology) have dictated an outcome not everyone is satisfied with?

What we have before us in the business of Egyptology bears some thought. We begin a process where we can direct what we as scholars need to have represent us on the academic publishing side of our field. What we come up with may not be pretty. It may have many warts and irregularities, but hopefully it is what we want it to be.

1. PRODUCTION ISSUES RELATED TO ARTICLE SUBMISSION (JULIE MASQUELIER-LOORIUS)

While some of the issues may be journal specific, they raise significant items that vary between journals and each may have a different solution.

(1) *Submissions must be homogenized by the copy editor.* Authors mainly submit their articles in various and ancient formats of word processors (WordPerfect, `.doc`, etc.), using different transliteration fonts (ThotM, ifaoTimes, WinGlyph), and hieroglyphic fonts (JSesh, MacScribe, WinGlyph, etc.). In some cases transliteration and hieroglyphic texts are simply hand written on hard copy. The copy editor loses too much time with this process. Generally authors may not know how to transform their texts in another format (.rtf for example).

(2) *Communication between the copy editor and the editor.* There are at time potential problems when passing on the fonts used in the *Revue d'Égyptologie* by the Editor (Peeters) to the copy editor.

(3) *The need for a `.pdf` version of an article.* Because of the various potential fonts, etc., authors need to submit a `.pdf` version of their submission to allow the copy editor to check all the fonts and contents of the papers, especially for transliteration, hieroglyphic texts, Greek texts and pictures.

(4) *Article submissions are done by post mailing.* Authors often send hard/paper copies plus CDs or DVDs. Fonts are given by the authors. The use of e-mailing for regular articles and some sort of ftp protocol would be easier. This method would translate into lower production costs, the use of less paper, and is much faster.

(5) *Copyright of pictures in the articles.* Authors often failed to obtain or are delayed in obtaining copyright permissions for the photographs used in their articles. Copyright may be difficult to obtain and this is especially true for cases of some objects preserved in museums. Then the photographs submitted are often of too low of a resolution to be reproduced.

Thus, the copy editor has become a corrector, a rewriter, a CAP (computer-aided production) technician, and a *de facto* mediator between the authors and the journal editor. In some cases the authors have specific *desiderata* that are not compatible with economical and technical constraints of the editor.

In conclusion we recommend that in order to obtain savings in time, money and paper, submitted articles must be formatted following specific standards. These would include a `.pdf` file which serves as a witness for contents, an `.rtf` file which can be easily read and retouched, pictures would be 300 dpi files in a suitable format such as `.psd`, `.tiff` or `.raw`, possibly `.ai` (illustrator data). If we ask people to choose for one transliteration font or hieroglyphic font or software, we should provide them macros that transform their files. The authors would receive by email or posted on the journal's Website specific formatting requirements, preferred file formats and instructions for submittal.

As you can see from these notes, the job of editing articles can be most complex because we have all been brought into the field from different directions, using different programs, and following approaches that sometime reflect variations found in different cultures, languages and academic schools.

2. ISSUES AND GOALS RELATED TO THE FUTURE OF JOURNAL PUBLISHING

The field of Egyptology was established in the 1800s with the translation of the Rosetta Stone. Following the publications of the "letters" of Champollion and others, the early scholars established a small group of journals to provide an outlet for the publication of research in the field. These early publications often contained the translations of texts and reports on the field work carried out in Egypt, and the opening salvos of the never ending discussions on the history and religion of the ancient Egyptians. For the most part (but not entirely) these journals were often representations of national interests in the field. Thus the German, the French, the British and the Italian publications dominated journal publications. Concomitant with that was the establishment of many archaeological institutes in Egypt or societies in Europe. These institutes and/or their sponsors also began producing publications including journals to publish the results of their work. Additional national institutes have developed over the years as additional countries have expanded or clarified their archaeological and academic presence in Egypt. Thus the journal I represent, the *Journal of the American Research Center in Egypt*, developed out of a relatively new institution doing work in Egypt (if one can call an institution only fifty years old as new).

But research institutes in Egypt have not been the only sources for Egyptological publications. Departments or institutes at universities have likewise fostered the development of publication sources and this has lead in recent years to a plethora of new journals representing expanded programs and groups. New journals from Spain, Australia and Armenia are just a few of the many new proceedings. As a field we have traditionally taken this as a positive sign of the growth of the field of Egyptology and have viewed the appearance of new journals as a good thing. (How can more not be good?)

We have also seen the development of several "second tier" or non-refereed journals such as *Göttinger Miszellen* (*GM*) and *Discussions in Egyptology* (*DE*). Here we had a method where scholars had short notes that they wanted relatively immediate feedback from the scholarly community. You sent in a short paper, the editors said yes or no, and if yes, they printed it with a dozen others similar items and the discussions began. In a sense Egyptology was at the forefront of blogging before the internet came around. I take note here that *GM* has expressly decided not to be seen as a non-refereed journal and has developed a referee system. This shows that journals can and do change their nature over time, if they have the will to do so.

Another publishing phenomenon for our field has expanded in the last decade and this is the plethora of memorial volumes dedicated to members of our field in honor of their contributions. These volumes are usually developed and edited by a student or close friend of the honored colleague. Originally in the field there were few of these *Festschriften* and they were remarkable volumes, the *Studies Griffiths* volume of 1932 being a classic example. Most *Festschriften* appeared as individual issues of a journal. As a scholar I now find myself besieged with invitations for *Festschriften* from a variety of sources. In some cases I have never even met these individuals and in other cases I have never even heard of the individual before (though that is rare). These *Festschriften* have been troubling to me as I often find those who are sponsoring them are "trolling the waters," hoping that they can get enough papers submitted to put together a really big volume to show how much that individual was appreciated. As a journal editor, I find the increase in the number of Festschrifts a troubling area of the field as these volumes for the most part are not peer reviewed and, while they do contain numerous significant studies, many times one must winnow through the chaff to find them.

So I have described a situation for some of the publishing activity found in the field of Egyptology. The one thing I have not described is one which is hard to define: what is the purpose of journal publishing in our field? Simply put we can say journal publishing exists for the sole purpose to expand our knowledge of the ancient Egyptians. If it were that simple, then I as an editor could simply retire now and let the blogs and tweets do all the work. So if it is not that, what is its purpose? One aspect is to provide a reasoned evaluation of a hypothesis offered by a member of the field. Since membership in the field is beyond the scope of this paper, there are some built in warts in the system. As an editor, a colleague sends me an article and it is evaluated to see if it will add anything new to the field. The process is similar for most journals: one or more colleagues evaluates the article and gives the editor feedback on whether the article should be published as it is, or should be published with revisions, or is to be rejected. The editor would then tell the author the verdict. That is the simple version.

Publish or perish is a standard by which most university academics live. Universities and their research institutes expect their scholars to publish material in refereed journals. They provide significant resources to some individuals for that purpose. They require a *quid pro quo*, so to speak, for their investments. Now getting an article published becomes an economic tool. Journals fulfill the role of evaluation of personnel as well as providing prestige for academic departments and institutes. "Our faculty has published XXX number of articles in refereed journals" is what many university promotions bellow. At the same time the various journals boast to members of their sponsoring societies that their journal has published articles by scholars around the world and therefore what may have started as a local Egyptology newsletter might now have the cachet of being an "internationally recognized" publication.

In a sense I am straying from the focus of this paper, but it is important to have a context for my coming statements. For our field is neither a young one nor a simple one. It is one that overlaps a number of theoretical approaches in the social sciences, humanities and natural sciences (if I may be so bold as to use the common American terminology). This adds to the confusion, for wherever our journals go they must reflect the varied academic environment they live in.

To increase the problems a minor phenomenon known as the World Wide Web has emerged in the last twenty years. The Web has literally changed everything. Using cutting edge, but simple technology, people have offered extensive amounts of raw data, reasoned commentary, pure speculation, and much drivel to the world. Does what we place on the Web count as a publication? Let me give an example of my own. A few years back I put on my Northern Arizona University Web site some of my research concerning the temple of Ghuieta in Kharga Oasis. Recently there appeared several articles by respected scholars who cited this research to support their theses. Now, my original intention was to publish the material from that Website as an article, but I never got around to it. Now it has been cited in the literature (peer reviewed) on several occasions. Should I now use that as a publication?

As you can see there is no simple answer to that question and that is the issue I want to raise with journal publishing. The World Wide Web has created a monster on the doorstep of academia and we as scholars need to take that monster and do something with it.

3. CONCLUSIONS AND PROPOSALS

Now I am not saying that we should not use online sources. On the contrary, my main goal is to state that in 15 years time, it would be simply foolish for our field to support hard copy journal publishing. I strongly believe that the field of Egyptology must adapt to the new technologies, but we must be very clear on what we are trying to accomplish when we do so. I propose that we adopt the following guidelines for the future of journal publishing:

(1) We in the field must adopt some computer aided mechanism for the production of hieroglyphs. Whether it is a version of Unicode or whatever, we need to be on the same page within our field. We cannot afford not to do it. I am not saying we must adopt any one system over another, but we must have an advanced system to print hieroglyphs. Adopting a single system would greatly aid the transportability of hieroglyphs over the Web. The Manuel de Codage is a first step, as were Winglyph, JSESH, and other programs. But we must have more.

(2) Related to that proposal we should have the "Informatique & Égyptologie" group make a recommendation on what we should do related to generating computer generated hieroglyphs and that should be presented to the International Association of Egyptology for consideration and forwarded to a newly formed online forum composed of the editors of the Egyptology journals. In that manner, direct feedback to and from publishers can be worked in quickly.

(3) Journals within the field of Egyptology should redefine their goals. All of the journals in the field should not try to be "the premier journal for all things Egyptological." I think that is not what we want for our journals. I do think we can make it clearer for all by having each journal demonstrate their major publishing focus. Thus I think journals such as *Bulletin de l'Institut Français d'Archéologie Orientale* (*BIFAO*), *Mitteilungen des Deutschen Archäologischen Instituts, Abteilung Kairo* (*MDAIK*), *Journal of the American Research Center in Egypt* (*JARCE*)

should focus on publishing works related to the efforts of their sponsoring organization. I did not say restrict, but focus. Other journals should clearly state that they are interested in specifically certain areas related to ancient Egypt, such as Egyptian history (thus the *Journal of Egyptian History*) or Egyptian language (thus *Lingua Aegyptia*). In many cases these journals may focus on areas related to the interests of their sponsors (a society or academic institution) and/or the Egyptologists that serve them.

(4) All current Egyptology journals should develop a plan to convert entirely to online formats within a ten year time frame. This plan would include the provision for the conversion of back issues to electronic format. Several current models exist. *BIFAO* has most of their back issues online (free access) available as `.pdf` files. *JARCE* uses the library journal service known as *JSTOR* where all back issues are available. It will be entirely up to each journal to determine how it will deliver its own journal online, whether by subscription, free or pay-per-view type of arrangement. A number of easily adaptable models exists currently and more may develop in the future. The ETANA Website (http://www.etana.org/) has a listing of journals with online versions of their material.

(5) I propose that copyright reside entirely with the author of the article and not with the journal or electronic publisher. Dissemination of the article would then be encouraged.

My proposals are not necessarily earth shattering. They do suggest a different approach to publishing, but one easily compatible with current practices. They are environmentally friendly as it will mean fewer trees will be cut down to provide for copies. Those who prefer to read hard copy can do so simply by printing off their own copies for personal use.

It is hoped that we will soon begin an on-going discussion group amongst the editors of Egyptological journals in order to make it easier for both authors and editors to deal with the myriad of issues related to publishing journal articles in the era of computers.

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.

Collection *Ægyptiaca Leodiensia*

La collection *Ægyptiaca Leodiensia* 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.

Collection *Ægyptiaca Leodiensia*

1. Jean WINAND, *Le Voyage d'Ounamon. Index verborum, concordance, relevés grammaticaux*, 1987.
2. Jean WINAND, *Études de néo-égyptien, 1. La morphologie verbale*, 1992.
3. Pierre KOEMOTH, *Osiris et les arbres. Contribution à l'étude des arbres sacrés de l'Égypte ancienne*, 1994.
4. Juan Carlos MORENO GARCÍA, *Études sur l'administration, le pouvoir et l'idéologie en Égypte, de l'Ancien au Moyen Empire*, 1997.
5. Dimitri LABOURY, *La statuaire de Thoutmosis III. Essai d'interprétation d'un portrait royal dans son contexte historique*, 1998.
6. Michel MALAISE & Jean WINAND, *Grammaire raisonnée de l'égyptien classique*, 1999.
7. Laurent BRICAULT, *Isis, Dame des flots*, 2006.
8. Jean WINAND & Alessandro STELLA (avec la collaboration de Laurence NEVEN), *Lexique du Moyen Égyptien, avec une introduction grammaticale et une liste des mots présentés selon le classificateur sémantique*, 2013.
9. Stéphane POLIS & Jean WINAND (avec la collaboration de Todd GILLEN), *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*, 2013.

Achevé d'imprimer sur les presses de Snel
Z.I. des Hauts-Sarts – Zone 3
Rue Fond des Fourches 21 – B-4041 Vottem (Herstal)
Janvier 2013