COLLECTION
ÆGYPTIACA LEODIENSIA

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

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

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

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

PRESSES UNIVERSITAIRES DE LIÈGE

ISBN : 978-2-87562-016-3

ÆGYPTIACA LEODIENSIA

9

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

COLLECTION
ÆGYPTIACA LEODIENSIA

Texts, Languages & Information Technology in Egyptology

Stéphane POLIS — Jean WINAND
With the collaboration of Todd GILLEN

Presses Universitaires de Liège

Texts, Languages & Information

Technology in Egyptology

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

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

Imprimé en Belgique

Collection *Ægyptiaca Leodiensia* 9

Texts, Languages & Information Technology in Egyptology

Selected papers from the meeting of the Computer Working Group
of the International Association of Egyptologists
(Informatique & Égyptologie), Liège, 6-8 July 2010

Stéphane POLIS & Jean WINAND (eds.)

With the collaboration of Todd GILLEN

Presses Universitaires de Liège
2013

Table of Contents

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.

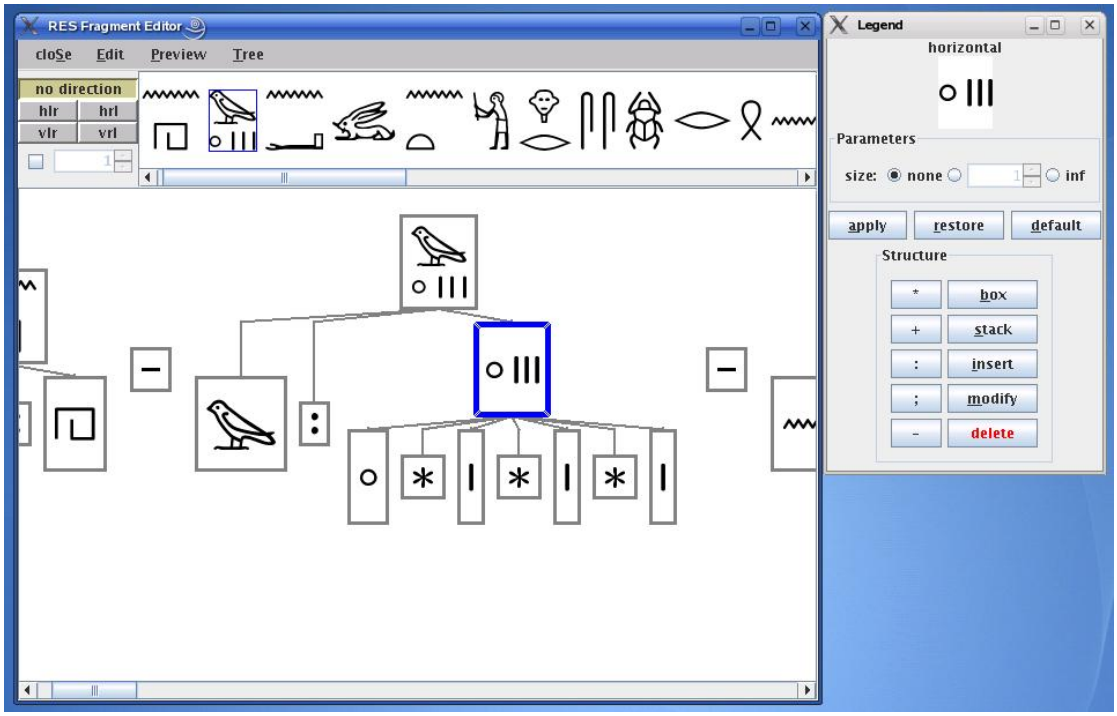

Figure 1. Graphical editor for RES[6]

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>
```

jbs3 jnbj mnw
wild mint,[5] hedge plants, pigeons,

[5] Meaning is uncertain.

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.

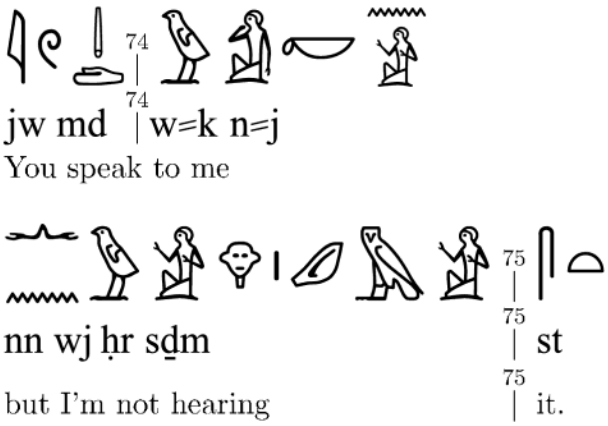

Figure 5. Interlinear text obtained from two independently created resources

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"/>
```

wild mint, hedge plants,

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.

Abstracts

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Carlos GRACIA ZAMACONA, A Database for the Coffin Texts

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

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

Azza EZZAT, The Digital Library of Inscriptions and Calligraphies

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

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

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