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

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.).

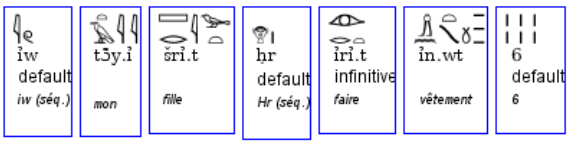

Figure 1. A sentence in Ramses' TextEditor

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.

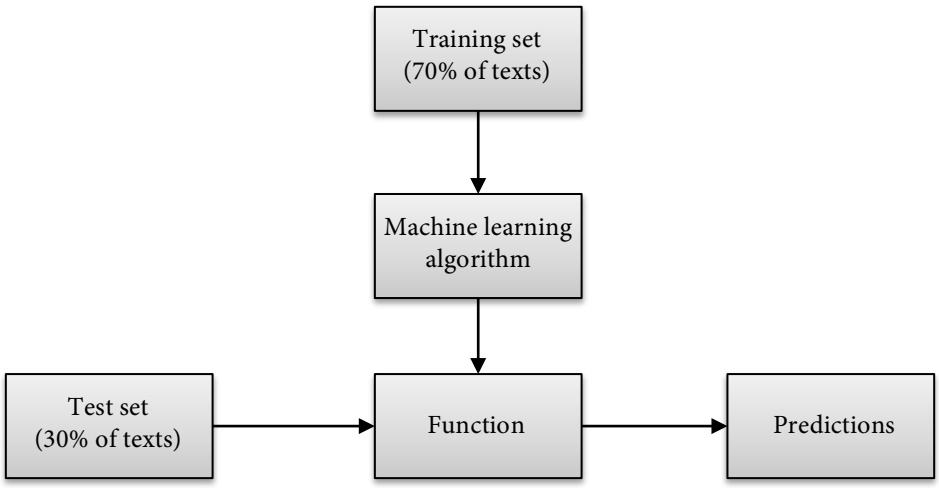

Figure 2. The learning and testing phases

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 B3s3* "does it belong to Bes?"

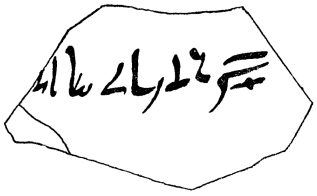

Figure 5. A typical oracular question (O. IFAO 866)

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.

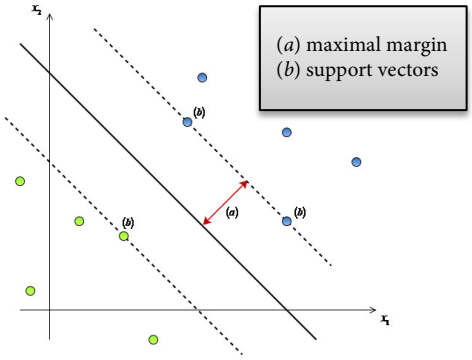

Figure 7. The principle of SVM

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.

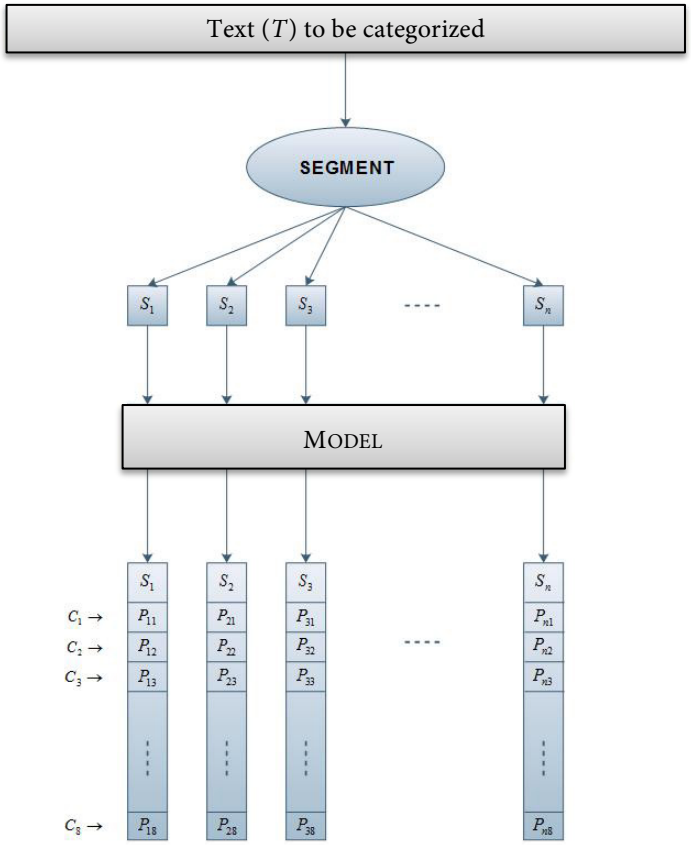

Figure 9. The Segment & Combine method (1)

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.

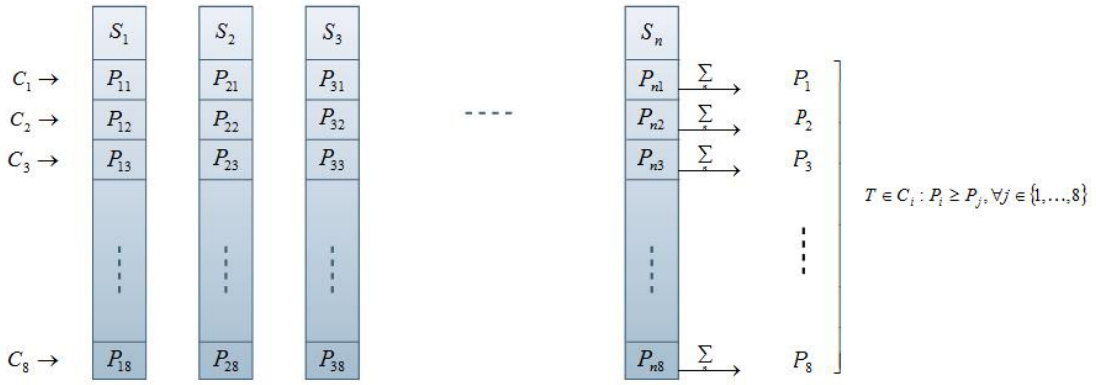

Figure 10. The Segment & Combine method (2)

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.

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.