



**HAL**  
open science

## Processing a Mayan Corpus for Enhancing our Knowledge of Ancient Scripts

Bruno Delprat, Mohamed Hallab, Martine Cadot, Alain Lelu

► **To cite this version:**

Bruno Delprat, Mohamed Hallab, Martine Cadot, Alain Lelu. Processing a Mayan Corpus for Enhancing our Knowledge of Ancient Scripts. 4th International Conference on Information Systems and Economic Intelligence - SIIE'2011, Feb 2011, Marrakech, Morocco. pp.198-208. hal-00577958

**HAL Id: hal-00577958**

**<https://hal.science/hal-00577958>**

Submitted on 24 Nov 2011

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Processing a Mayan Corpus for Enhancing our Knowledge of Ancient Scripts

Bruno Delprat<sup>\*†</sup>, Mohamed Hallab<sup>‡§</sup>, Martine Cadot<sup>¶||</sup> and Alain Lelu<sup>||\*\*,††</sup>

<sup>\*</sup> INALCO, École doctorale - 49bis, avenue de la Belle Gabrielle, 75012 Paris, France

<sup>†</sup> Centre d'Études des Langues Indigènes d'Amérique, SeDyL-CNRS - 7, rue Guy Môquet, 94801 Villejuif cedex, France  
Email: brunodelprat@club-internet.fr

<sup>‡</sup> CNAM, La Défense, 15 Ave d'Alsace, 92400 Courbevoie, France

<sup>§</sup> Université du 7 Novembre, École Supérieure de Technologie et d'Informatique  
4, rue des Entrepreneurs Charguia II - 2035 Tunis-Carthage, Tunisia  
Email: mohamed.hallab@yahoo.fr

<sup>¶</sup> Computer Department, Université Nancy I, France

<sup>||</sup> LORIA, Campus Scientifique - BP 239 - 54506 Vandœuvre-lès-Nancy Cedex, France  
Email: martine.cadot@loria.fr

<sup>\*\*</sup> Université de Franche-Comté, Laseldi, 25030 Besançon Cedex, France

<sup>††</sup> Institut des sciences de la communication du CNRS, 20 rue Berbier-du-Mets, 75013 Paris, France  
Email: alain.lelu@univ-fcomte.fr

**Abstract**—The ancient Maya writing comprises more than 500 signs, either syllabic or semantic, and is largely deciphered, with a variable degree of reliability. We applied to the Dresden Codex, one of the only three manuscripts that reached us, encoded for  $\LaTeX$  with the  $\mayaTeX$  package, our graded representation method of hybrid non-supervised learning, intermediate between clustering and oblique factor analysis, and following Hellinger metrics, in order to obtain a nuanced image of themes dealt with: the statistical entities are the 214 codex segments, and their attributes are the 1687 extracted bigrams of signs. For comparison, we introduced in this approach an exogenous element, i.e. the splitting of the composed signs into their elements, for a finer elicitation of the contents. The results are visualized as a set of “thematic concordances”: for each homogeneous semantic context, the most salient bigrams or sequences of bigrams are displayed in their textual environment, which sheds a new light on the meaning of some little understood glyphs, placing them in clearly understandable contexts.

## I. INTRODUCTION AND PROBLEMATICS

The logo-syllabic writing of the ancient Mayas, in use during more than 13 centuries, reached us through rich inscriptions on monuments, ceramics and three divinatory almanacs, which constitute nevertheless a small volume of available texts: three manuscripts and about a thousand short inscriptions.

The objective of the work here presented is to retrieve the main semantic contexts of glyphs usage, in the spirit of [1] approach to semantic categories, in order to bring together in a common context deciphered glyphs, and those which are less or not at all understood. On the long run, it could result in providing the mayanist scientific community with such a contextualization for the whole available Maya corpus.

We chose for this first computer encoded corpus a representation method characterized as follows:

- 1) unsupervised, so that from co-occurrences of glyphs, or other text attributes, arise interpretable contexts comparable to what is already known,
- 2) fuzzy: to every statistical unit, elementary text or attribute, is allocated a centrality value which strength varies with the various uncovered contexts. Accordingly, elements with a weak representativeness in the analysis will show low values in all contexts; polysemous or syntactic elements will be central in several contexts, i.e. show a strong value in several clusters; elements with a univocal signification will be central in a single cluster only<sup>1</sup>,
- 3) compatible with the relative scarcity of available sources, by contrast with the statistical approach to language models that require to collate millions of occurrences [3].

## II. PRINCIPLES OF MAYA WRITING AND ITS ENCODING FOR $\LaTeX$

The elementary sign of writing is the glyph. One or more glyphs are assembled together to harmoniously fill the preset rectangular space of a *cartouche*. Maya manuscripts texts are organized in blocks of 2 to 16 cartouches which constitute as many sentences, often followed by associated numbers, dates and possibly a picture. Depending on the number of cartouche slots available on the almanac page for writing a short or long sentence, the scribe would squeeze in or spread out writing signs among the cartouches to avoid empty boxes and obtain a nice looking page layout.

<sup>1</sup>In this respect our methodology differs from “fuzzy clustering” [2], because the sum of the centralities of a statistical unit in the different clusters is not forced to 1; here centrality does not convey the uncertainty of the belonging to a cluster (a probability), but a contribution to context building.

### A. General principles of Maya writing

Through the analysis of the Maya codices of Dresden, Madrid and Paris, we determined that a complete glyphic cartouche is made of 1 to 5 basic signs or glyphs, from a catalogue of 509 glyphs. Affixes, like 031  *ni*, present rotation patterns and symmetries, following a determined rule, around the central elements such as 204  which orientation is fixed. Affixes are generally syllabic value signs which can combine together or with a central element to write a Maya word. Central elements are generally of logographic nature, corresponding to a morpheme or a word, and read globally like 204  *KIN* sun, day. The general orientation rule of the affixes is the following:

(063) *te*      
 For example :    

A complete glyphic cartouche often corresponds to a lexical entry with preceding and following grammatical affixes as in 204.031  *KIN-ni* sun, day, but it can also in some cases correspond to two words if they are short, or more rarely otherwise to a part of an expression spelt over two cartouches.

### B. Maya glyphs computer input and composition with *mayaTeX*

In the early 1960s' Soviet cryptologists and historians [4] undertook the first computer encoding of the cursive maya writing in a catalogue of about 500 glyph forms, then coded on magnetic tape the Dresden and Madrid codex to process the texts with the computer possibilities of that time, a pioneering work almost completely forgotten since.

A prerequisite to the computer encoding of Maya texts is the structural analysis of the written form of the codices, which includes the formulation of basic hieroglyphic signs composition rules within the Maya cartouche and the definition of a graphic grammar, work done as part of an ongoing doctoral dissertation [5].

*MayaTeX* [6], an original computer tool for the input and edition of Maya hieroglyphic texts developed under *TeX*, is used for the composition of the palaeography and the dictionary included in the dissertation. The two main glyph composition operators within a Maya cartouche are the point “.” which associates two elements 117  and 260  by left-right juxtaposing 117.260 , and the slash “/” which places an element 400  above another one 010  to yield 400 / 010.030 . Operators “.” or “/” are also active on a subset of glyphs in brackets (154.123) , to yield (154.123)/177 .

*Ligatures*: One or several affixes or central elements can be melted inside a geometric central element or, more frequently, inside a head variant instead of being simply attached to it, forming a ligature as a single bound form. Ligature glyphs appear in the catalogue as graphic elements with specific codes 373  *Cacau* D7c (2), which in fact decomposes into simpler glyphs: 369 <023/023>. The operator < > indicates that both affixes 023  are placed in the centre of 369 . Within Maya texts, both forms: melted as a ligature (described by one code), and separately drawn (described by 2 - 3 codes) are equivalent and constitute graphic variants. Our catalogue includes 108 ligature glyphs.

### III. THE DRESDEN CODEX AND ITS DECIPHERMENT

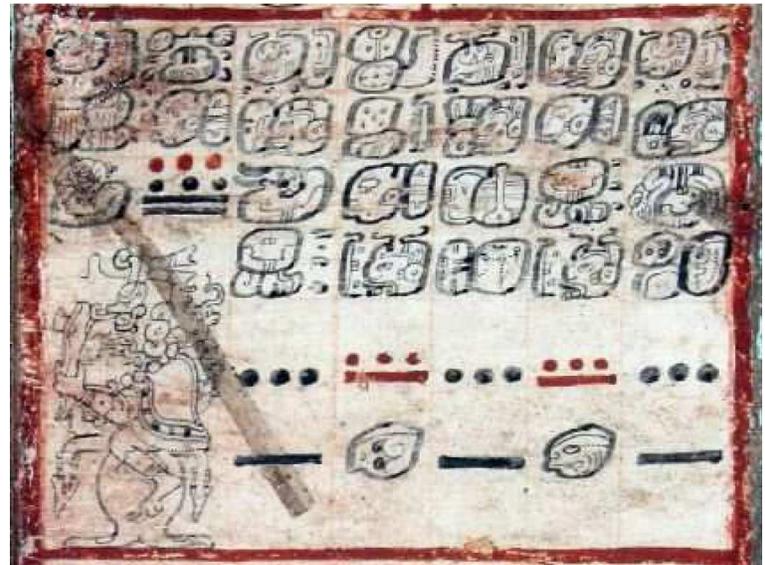


Figure 1. Dresden codex Segment 30b

*A divinatory almanac*: The Dresden codex, one of the three only handwritten Mayan texts which reached us and dating probably from the 15th century, is constituted of all 76 almanacs of 5 main types: divinatory almanacs of *Tzolkin* calendar of 260 days dedicated to various divinities, prophecies of *Haab* solar year of 360 days plus 5 extra leap days, and of *Katuns* or cycles of 52 years, almanacs of four cardinal directions dedicated in Chac, the God of the water, astronomic tables such as the stages of Venus, solar eclipses and lunar eclipses, and almanacs of the formalities of the New Year and of the flood associated with *Katun* cycle of 52 years. These texts are in general independent one another and not the successive chapters of a Western book to be read from the beginning to the end.

Figure 1 reproduces the middle segment, numbered 30b, of folio 30 with the corresponding hieroglyphical palaeography of the 2nd text block from the left encoded under

maya $\text{\TeX}$ . This almanac is part of a series dedicated to the tasks performed by Chac, the God of rain and water, who appears in the form of the avatars of the four directions. Corresponding offerings are a dish of meat and a particular colour tribute for each of the four directions:

				
400/010.030	+176/204.031	117.260	133/111.023	423/515
<i>tseI-ah</i>	<i>lakin</i>	<i>chac-xib</i>	<i>kabil</i>	<i>cehel-uah;</i>
Was standing	East	red man	sweet	deer tamal
				
530.112	515/504.013	026.401		
<i>Chac</i>	<i>hanal</i>	<i>u-bool?</i> ( <i>u-can?</i> )		
god Chac	meal	its tributes		

*God Chac as a red man stood on the East; its tributes are a meal of sweets and deer tamales.*

			
903	905	808	710
3	5	8	<i>Oc</i>
3x20	+5	8	<i>Oc</i>

65 [days up to] 8 *Oc*.

*The text corpus:* The 74 codex folios, numbered by [7] and encoded following the catalogue of [4] with some complements for maya $\text{\TeX}$ , do not read in a linear way the one after another. Indeed, every divinatory almanac or text with astronomic table of the codex is painted across several folios, for example the upper segments of folios 4a, 5a, 6a and 10a. The text statistical units considered in this research are the 214 folio segments as defined by the Maya scribes, generally separated by a red line, which usually correspond to upper, middle or lower part of folios.

Each folio segment includes 1 to 8 blocks of hieroglyphic cartouches, one block holding for one sentence. The distribution of the number of sentences per folio segment is concentrated: 77% of the segments include 2 to 4 sentences (see Figure 2).

From a statistical point of view, the Dresden codex corpus is composed of 9938 glyphs including dates and numbers, or 7549 proper text glyphs, composed of 409 individual writing signs, among which 345 textual elementary or ligature glyphs. Individual glyphs occur in the codex with 1 to 252 repetitions<sup>2</sup>. Unsurprisingly, the distribution of the glyphs follows a Zipfian allure (see Figure 3), linear in log-log coordinates and typical of a power law, common to all corpuses from any language origin, confirming that Maya glyphs write human language. The size of the text segments, measured by the glyph number, follows a distribution law

<sup>2</sup>In fact, the most represented sign with 411 repetitions is  replacing an erased text glyph. Syllabogram affixes 111  *li* and 026  *u-* (his, her, 3rd person relative pronoun) are the most frequent textual glyphs both with 345 occurrences.

with a binomial allure (see Figure 4), as text segments in many other corpuses do.

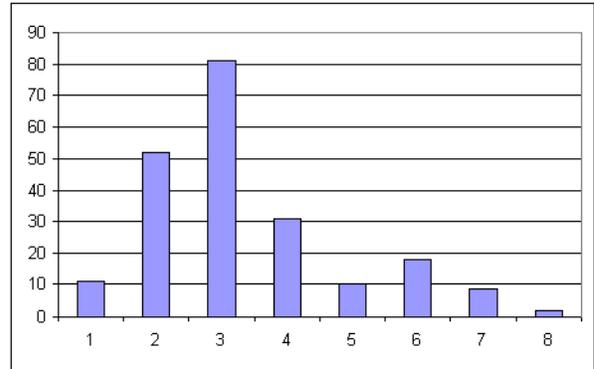


Figure 2. Characterizing folio segments in the Dresde codex; x-axis, number of sentences in a segment; y-axis, number of segments. For example, one may read: 81 segments consist of 3 sentences

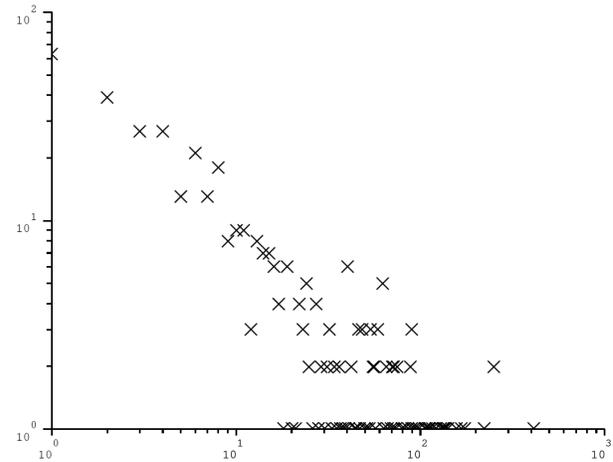


Figure 3. Recurrence of glyphs in the corpus, measured without breaking out ligatures. Coordinates are log-log: x-axis, occurrences of a glyph in the corpus; y-axis, number of glyphs presenting these occurrences. For example, one may read: there are 39 glyphs occurring 2 times.

Each sentence includes 1 to 26 cartouches, including numbers and dates. The above-mentioned aesthetic requirements account for the fact that more than one sentence upon three consists of exactly 6 boxes. It follows also that about one segment upon four is composed of three sentences made of 6 boxes. The corresponding distributions are thus very asymmetric, with a strong modal value, respectively 6 boxes per sentence and 18 boxes per segment.

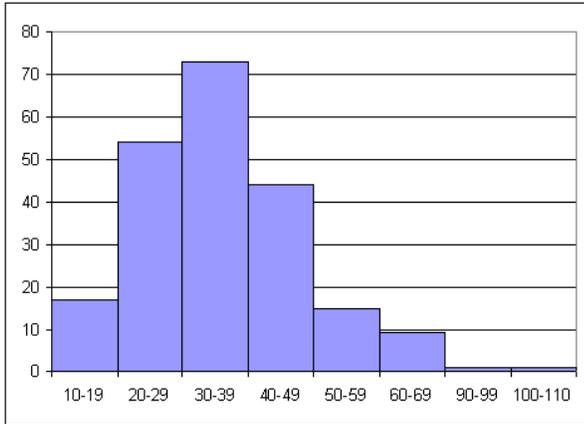


Figure 4. Characterizing folio segments in the Dresden codex – Histogram of the number of glyphs per corpus segment: x-axis, number of individual glyphs in each segment; y-axis, number of corresponding segments in the corpus. For example, one may read: 73 folio segments consist of 30 to 39 individual glyphs.

*The decipherment:* It benefits from favourable factors: the Mayan languages are still spoken at present, and we have at our disposal the prophecies of the *Chilam Balams* [8], which are partly the Yucatec transcription in Latin alphabet of divinatory texts similar to the ones found in the three remaining hieroglyphical divinatory codices. The meaning of the logo-syllabic signs, i.e. glyphs, is established with certainty for more than a fifth, and with reasonable credibility for more than a half.

They fall into three main types. The **logogram**, or morphemic sign with both semantic and phonetic values, most often monosyllabic of the consonant-vocal-consonant (CVC) form, as glyph 204  representing a four petals flower, symbol of the sun and with phonetic value *KIN* sun, day. The **syllabogram**, noting a CV or VC type syllable, as glyph +176  with phonetic value *la* which is used to write different words combined with another syllabogram or a logogram as +176/204  *la-KIN* (East). Syllabograms are derived from CV(C) value logograms in which the second consonant is weak. The **phonetic complement**, a sign of phonetic value CV, VC or a part of the C(V) value of the logogram to which it is associated, as in the word 204.031 , where 031  is the phonetic complement *ni* or *in*, postfixed to *KIN* (the sun, day) to confirm the reading spelling. A word can be written through different combinations of glyphs that note the same phonetic value, as for *lakin* (East) indifferently written by allographs +176/204 , 176/204.031 , 204/176.031  or 031.+176/204 . In spite of this constituent commutability, one may notice a large majority of frozen textual cartouche arrangements: as the whole corpus comprises 3034 cartouches of text with 1258 unique occurrences, 1197 of

them are composed of glyphs that strictly show no variation of their graphical layout, which amounts to 79% of the occurrences. In the 61 remaining occurrences, the glyphs may be turned around or permuted. It is noticeable that no more than 19 of the latter give rise to variants, in the coding of which the ordering of the basic (i.e. non-positional) codes may differ: e.g. the three variants 032.063/153 , 032.-063/153 , 032.153/063  occur respectively 6, 1, and 3 times.

#### IV. THE PROCESSING CHAIN

The corpus encoded under *mayaTeX* was first pre-processed to extract n-grams of individual glyphs. Then, a further described clustering algorithm, programmed with *Scilab*, provides for each cluster the ordered and valued lists of n-grams and documents (i.e. folio segments, see below) that characterize the cluster. In a post-processing stage, these lists are improved appending their context cartouches, and visualized using *mayaTeX* to enable their linguistic interpretation.

One may oppose that the glyph-codes ordering induced by the chosen coding scheme is fairly conventional, and that it should be better to use, instead of bigrams, unordered glyph “2-itemsets” occurring within a cartouche, whatever their graphical placing: the dominant frozen configuration reported at the end of section 3 shows that such a coding scheme would be needlessly redundant, compared to a simple bigram coding rule. Of course this freezing may be due to the sole writing habit of one putative scrivener: the coding of the entire Maya known corpus (the three codices and the diverse other written material) may necessitate such an extra coding convention.

##### A. Pre-processing

*Choice for the split into text units:* How to split the corpus into text statistical units – the comparison of which is the rationale for the analysis – directly impacts the granularity of the analysis: too fine a split, for example at cartouche or glyph level, would favour syntactic elements, not our present objective; a rougher split would privilege semantic elements, but the risk then is ending up with too few elements to obtain a clustering at the wished fineness level. The in-between solution adopted here is to consider folio segments as defined by the Maya scribes, knowing that the texts in these segments can be spread out on several folios. The number of page segments is a priori compatible with a granularity of analysis needed for around ten semantic clusters, allowing to go beyond obvious and known text divisions (stages of Venus, prediction of eclipses, etc.). A finer division, at the level of the Maya sentence was also tested, to validate our splitting up choice, and to extend the analysis onto the syntactic plan.

Excerpt from text segment 5b:

 ...  ⇒ Maya sentence  
 ⇒ cartouche     ⇒ glyph;     ⇒ glyph  
 ⇒ cartouche

*Why n-grams ? Choice of n:* For the experiments presented here we decided to characterise each portion of text by its vector of glyph bigram frequencies, as they appear from maya $\text{\TeX}$  coding: observed combinations of these bigrams, considerably lesser than  $500^2$ , can be easily handled with current computer technology without needing to recur to a H-coding compression of the number of codes, as we did in the context of other applications in the past [9]; our present choice is not to have bigrams crossing cartouche borders, as cartouches correspond most frequently to distinct phrases. The use of bigrams constitutes a flexible and minimal way of reflecting text sequentality. Maya bigrams correspond most often to a part of the 3 to 5 signs cartouche, and therefore to words or phrases which are displayed after each bigram in the tables.

Our software for presenting the results post-processes the (coded) outputs of the clustering stage, and was adapted to the specificities of the Maya writing. One of the difficulties we had to solve was the following: glyph codes are surrounded by position codes (above/below, sign symmetry or rotation, ...) that are easy to filter to obtain glyph bigrams free from this information. The other way round, it is necessary to restore these elements to display all the salient bigrams in a cluster in their original graphic context: we chose to provide the mayanist user, for each important bigram in a cluster, with all different cartouches to which it participates, establishing a kind of a concordance list at the level of each cluster.

*Extraction of n-grams:* Original texts are transformed by reduction of position characters  $/:<$  and filtering orientation attributes  $()*?-!+@|>$ . A window of length 7 (2 Maya glyphs separated by a point) moves along the text, 4 characters at the same time. A blank space stops current scrolling by the  $N=2$  glyphs window, then re-initializes it.

#### Example text

Palaeography:			
Coded:	990.172/056	(154.123)/306	*002c
Transformed:	990.172.056	154.123.306	002c
Extracted bi-grams :	990.172, 172.056,	154.123, 123.306	

#### B. The unsupervised learning process

*Distributional distance and cosine:* A series of ancient works [10], [11], [12], [13] focused on what some authors coin as *distributional distance*, and others *Hellinger distance*: it deals with the usual Euclidean distance (equally weighted dimensions), between two points  $t_1$  and  $t_2$ , the

coordinates of which are given by the vectors  $\mathbf{z}_{t_1}$  and  $\mathbf{z}_{t_2}$ , at the *surface of the unit hypersphere* in the space of the I descriptors. These points depict text segments, defined by the following transformation of the original frequency count data:

$$\mathbf{z}_{t_1} : \left\{ \sqrt{\frac{x_{it_1}}{x_{.t_1}}} \right\} \quad ; \quad \mathbf{z}_{t_2} : \left\{ \sqrt{\frac{x_{it_2}}{x_{.t_2}}} \right\} \quad (1)$$

where  $x_{it}$  is the frequency count of descriptor  $i$  in segment  $t$ , and  $x_{.t}$  is the total number of descriptors in segment  $t$ .

The distributional distance  $Dd(t_1, t_2)$  between the text segments  $t_1$  and  $t_2$  yields:

$$Dd(t_1, t_2) = \|\mathbf{z}_{t_1} - \mathbf{z}_{t_2}\| \quad (2)$$

where  $\|\mathbf{x}\|$  stands for the Euclidean norm of vector  $\mathbf{x}$ .

This distance is the length of the chord associated with the  $(\mathbf{z}_{t_1}, \mathbf{z}_{t_2})$  angle - it equals 2 when the vectors are opposite,  $\sqrt{2}$  when they are orthogonal. This distance may seem trivial and somehow arbitrary, as one may question why such an unusual normalization instead of the usual colinear one  $\left\{ \frac{x_{it}}{\|\mathbf{x}_t\|} \right\}$ , but its properties are interesting:

- 1) Contrary to the  $\chi^2$  distance underlying Correspondence Analysis [14], it can take into account vectors with negative components, a useful property for “symmetric” codings such as *Yes / No / Dont know*, or oriented flow matrices – for economics, physics, ...
- 2) This distance is closely connected to Renyi’s order  $\frac{1}{2}$  information gain [15] provided by a distribution  $\mathbf{x}_q$  when the distribution  $\mathbf{x}_p$  is known:

$$I^{(1/2)}(\mathbf{x}_q/\mathbf{x}_p) = -2\log_2(\cos(\mathbf{z}_p, \mathbf{z}_q)) = -2\log_2\left(1 - \frac{Dd^2}{2}\right) \quad (3)$$

- 3) It is specially suited to “directional data” [16] vectors such as encountered in textual data, the angle of which are relevant, not their lengths,
- 4) When the  $\mathbf{z}_t$  vectors are sparse, as is the case for text data, computation time is strongly reduced,
- 5) Last but not least, [11], [12] have shown that it benefited from the same “distributional equivalence” property as the  $\chi^2$  distance used in Correspondence Analysis: when two descriptors with the same relative profile are merged, the distances between text segments are unaltered. In other words, this property ensures the stability of the distance system when elements with similar distributions are merged or split, whatever column-wise or row-wise.

*Our unsupervised clustering method:* The principles at work in the axial KMeans clustering method [17] are:

- 1) turning the documents raw data cloud, i.e. data cloud of the text segments, into a normalized data cloud at

the surface of the unit hypersphere by means of the (non-colinear) mapping:

$$x_{ij} \rightarrow \sqrt{\frac{x_{ij}}{x_i}} \quad (4)$$

where  $x_i$  is the sum of the  $i$ th document-vector  $\mathbf{x}_i$ ,

- 2) splitting this cloud into  $K$  sub-clouds, each one provided with an axoid issued from the centre of the hypersphere. Each point belongs then to the cluster of its maximum projection on all axoids.

The axoid of each sub-cloud is defined as the first component extracted by Spherical Factor Analysis (SFA) [12] (“difference to the null data table” option): in matrix notation, if  $\mathbf{X}$  is the (document  $\times$  attribute) data table, summing to  $x_i$  and  $x_j$  row-wise and column wise respectively, if  $\mathbf{D}_r^{-\frac{1}{2}}$  is the diagonal matrix of  $\left\{x_i^{-\frac{1}{2}}\right\}$  (respectively  $\mathbf{D}_c^{-\frac{1}{2}}$  with  $\left\{x_j^{-\frac{1}{2}}\right\}$ ), the singular value decomposition (SVD) of:

$$\mathbf{X}^{\frac{1}{2}} = \left\{x_{ij}^{\frac{1}{2}}\right\} \quad \text{writes :} \quad \mathbf{X}^{\frac{1}{2}} = \mathbf{U}\mathbf{D}\mathbf{V}' \quad (5)$$

where the prime sign indicates the transposition of a matrix, and where the orthonormality conditions:

$$\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I} \quad \text{hold.} \quad (6)$$

The SFA factors write:

$$\mathbf{F} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}\mathbf{D} \quad (7)$$

$$\mathbf{G} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}\mathbf{D} \quad (8)$$

The reconstitution of the data writes:

$$\mathbf{X}^{\frac{1}{2}} = \mathbf{D}_r^{\frac{1}{2}}\mathbf{F}\mathbf{D}^{-1}\mathbf{G}'\mathbf{D}_c^{\frac{1}{2}} \quad (9)$$

Note that this process is formally related to correspondence analysis (CA), where SVD applies to the transformed matrix:

$$\mathbf{Q} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{X}\mathbf{D}_c^{-\frac{1}{2}} \quad (10)$$

leading to:

$$\mathbf{Q} = \mathbf{U}_{ca}\mathbf{D}_{ca}\mathbf{V}'_{ca} \quad (11)$$

and where the CA factors write:

$$\mathbf{F}_{ca} = x_{..}^{\frac{1}{2}}\mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}_{ca}\mathbf{D}_{ca} \quad (12)$$

$$\mathbf{G}_{ca} = x_{..}^{\frac{1}{2}}\mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}_{ca}\mathbf{D}_{ca} \quad (13)$$

Eventually, the reconstitution writes:

$$\mathbf{X} = \frac{1}{x_{..}}\mathbf{D}_r\mathbf{F}_{ca}\mathbf{D}_{ca}^{-1}\mathbf{G}'_{ca}\mathbf{D}_c \quad (14)$$

Geometrically speaking, CA maps the raw data cloud onto the surface of a “stretched simplex” [18] whose barycentre is pointed out by the first factor, a trivial vector of ones (the corresponding first eigenvalue is equal to one).

This contrasts with the SFA, where the document factors, i.e. projections of the documents onto the first axis, define the centrality indices of these documents<sup>3</sup>, and the squared first eigenvalue  $\lambda_1^2$  defines the portion of the data table sum  $x_{..}$  accounted for by the first-order reconstitution of:

$$\mathbf{X} \simeq \left\{x_{ij} = \frac{1}{\lambda_1}\mathbf{F}_1^2(i)\mathbf{G}_1^2(j)x_i x_j\right\} \quad (15)$$

where  $\mathbf{F}_1(i)$  stands for the  $i$ -th component of the first row-factor, and symmetrically for  $\mathbf{G}_1(j)$ .

In our specific clustering application, the sum over all clusters of their squared first eigenvalue accounts for 25.21% of the data.

In the perspective of our present analysis, an important property is that of duality: to the principal axis of the document data cloud, expressed by a vector with one coordinate per attribute, corresponds the formally symmetric principal axis of the attribute data cloud, expressed by a vector with one coordinate per document (“transition formula”):

$$\mathbf{F}_1 = \frac{1}{\lambda_1}\mathbf{D}_r^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}}\mathbf{D}_c^{\frac{1}{2}}\mathbf{G}_1 \quad (16)$$

This approach allows one, starting from a crisp clustering of the text segments, to infer nuanced representations:

- 1) *typicality* of a text segment, relative to several semantic contexts, and not to a single one;
- 2) *specificity (cue-validity)* of each bigram in the diverse contexts;
- 3) dual relations between these indices, that do not exist to our knowledge with other crisp or fuzzy clustering methods.

### C. Post-processing : the presentation of results

As will be detailed below, our unsupervised classification algorithm provides the ordered and valued lists of bigrams and characteristic documents of this cluster. To be able to interpret the theme dealt with in a cluster, additionally to documents titles, we must have the list of its most salient words (in our case: cartouches) available. The construction of the list for a given cluster requires to perform a 2nd pass on the documents of this cluster. For this reason, we extract altogether the bigram and its corresponding cartouche (i.e. phrase delimited by 2 blank spaces and which sequential number in the original text corresponds to that of the current text unit).

<sup>3</sup>It follows straightfully from the properties of eigenvectors that the first eigenvector of the inter-row cosine table  $\mathbf{D}_r^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}}\mathbf{X}^{\frac{1}{2}}\mathbf{D}_r^{-\frac{1}{2}}$  is  $\mathbf{F}_1$ . This vector can be interpreted as the “eigenvector centralities” [19] of the nodes of the valued graph whose adjacency matrix is this cosine table.

Tables for the resulting clusters are presented by order of decreasing centrality of Maya phrases for each cluster. We gathered together bigrams corresponding to the same phrases.

## V. EXPERIENCES AND RESULTS

Several series of experiments were performed. We present here only two, having in common the same pre-processing and unsupervised classification parameters, but with different corpus encoding methods: first without splitting complex ligature glyphs into developed encodings, and secondly splitting ligature forms into simpler elementary glyphs. The corresponding ligature decomposition table represents the exogenous information introduced for the second experiment.

After several tests, the number of requested clusters (10) appeared as a reasonable balance between the number of analyzed texts segments (215) and expected analysis precision. bigrams are established from elementary glyphs found within each Maya text cartouche, ignoring links with adjacent cartouches. With this pre-processing, Maya text cartouches containing one single glyphic element are ignored. Furthermore, as our experimentation targets text within the Maya codex and not its astronomical or calendar tables, we excluded Maya cartouches containing nothing but red digits (in vigesimal notation) of series 8xx  .

...  or black digits of series 9xx  ... For example, the digits couple 808/917  was ignored.

On the contrary, combinations of a digit (series 8xx or 9xx) with a non numerical glyph are processed, including:

- dates of the 260 days *tzolkin* calendar round which combine a red digit from 1 to 13 with one of the 20 days in series 7xx  ...  (as for example 807.704  7 Kan),
- dates in the 365 days *haab* year calendar including maya month names (as for example 908.411/515  8 Cumku), or
- numbers associated with words in the text, that often correspond to names of deities (as for example 908.255/220 .

### A. Without exogenous information

Ligatures of elementary glyphs were not broken down into their constitutive elements and appear as individual glyphs. Table II presents a 46 bigram cluster in which 21 phrases or hieroglyphical words, coloured in red within cartouches in the table of Maya phrases and bigrams of the cluster. We present here the transcription, translation and corresponding segments of the corpus, ordered by centrality ranking and in the present state of Maya writing decipherment. Only the 14

phrases corresponding to several occurrences in the corpus are displayed.

We note that, linked to the foreground in this cluster, the Mayan terms of the cycles of Mayan long count are standing out, with *Kin* 1 day, *Uinic* 20 days, *Tun / haab* 360 days or 1 year, *Katun* 20 years, *Baktun / pictun* 400 years. The origin date 4 Ahau 8 Cumku , from Mayan long count, stands out less prominently. Different parts of the almanac of the Number snake in long count and table of the numerals of 91 days constitute the basics of this cluster (6 documents among 10 in the cluster), but half of the sections of this almanac is more or less missing there. Sections 61AB and 69A of the same structure, with very similar and parallel texts (dates in long count) are rather closely linked (almanac of the Number snake there considers long count and so does the table of the numerals of 91 days). A strong centrality is given to bigrams of sections 61AB and 69A, introducing this single case in the corpus, at the expense of other elements in the cluster. The section 31b of the almanac of mythical and historical dates is a summed up resumption of the previous almanac, and its inclusion in the cluster 1 is relevant.

Besides, we notice that the thematic sorting of the text segments in the clusters coincides well with their distribution in the different almanacs of the codex.

### B. With exogenous information : splitting ligature glyphs into developed encodings

For a significant number of glyph forms (107 upon the 402 signs, except the figures, of the Mayan font), considering their global inclusion in a code, their link with their essential components is lost to our method of unsupervised classification, while it appears however visually straightaway. See for example: the global coding 455 , and developed coding 454/111 , which corresponds in fact to the fusion of two stacked glyphs  of linked bigram 454.111.

As for the previous experience, but in a more clear-cut way, the 10 required clusters bring to light with strong centrality terms carrying related notions, as does the Table III for cluster 5. The emergence of content clusters is made more specific by the decomposition of tying, which enriches the number of contexts and cases of glyphs which are not very frequent in the corpus (typically with less than 10 occurrences). Also, it allows to categorise better, in the context of the corpus, the grammatical affixes of the significant glyphs, when an affix as 030  -ah (the mark of the accomplished), is written merged with the verb which it concerns in a tying 374 / 318.030  or 318.030 , instead of being juxtaposed separately as  or .

## VI. RELATED APPROACHES

Besides the pioneering work of [4], we are not aware of any data mining studies of Mayan texts other than ours. However, data mining approaches to Chinese texts present both strong similarities and strong differences:

- Considered at the graphical and structural levels, each Chinese character can be portrayed as a "cartouche", i.e. a 2-dimensional display of semantic or phonetic components, among which 214 semantic "keys", and/or as a composition or repetition of other characters, resulting in a complex description [20], not unrelated to the above-described compositional aspects of Mayan cartouches: these components are distorted so as to fit the entire Chinese character into a square.

- These "cartouches" have been historically frozen into a large, but finite, number of well-specified characters: a few ten thousand characters listed in usual dictionaries and now computer-encoded, and a long distribution tail of rare ones, including hapax and variants, summing up to about 100 000.

- As a consequence, and given the lack of explicit separation marks between words, most Chinese search engines and many data mining studies, such as [21], [22], [23] use character n-grams as elementary features for coding Chinese texts, mainly 1-grams to 3-grams.

- These studies use, as we do, a vector count of n-grams for computing similarities between texts. The big difference is that they can use immense corpuses of many ten thousand news articles or Web pages, compared to the restricted exhaustive Mayan Corpus of three codices and a few thousand ceramic and monumental inscriptions.

- [23] builds his supervised categorization of Chinese Web pages on a variant of the unsupervised Latent Dirichlet Allocation clustering method [24], which shares the same drawbacks of our Kmeans-based method: the process is initialization-dependant, and the number K of clusters has to be specified.

## VII. CONCLUSION AND PERSPECTIVES

The coding of the Dresden codex under  $\text{mayaTeX}$  permitted to initiate a series of experiments highlighting semantic clusters through non-supervised learning. A cluster analysis giving nuanced results, thanks to our original method, was performed, based on the text division in page segments, characterized by their Maya bigram profile. This analysis confirmed the grouping of well understood glyphs into already known semantic clusters, and for other glyphs, more subject to controversy, permitted to orient their interpretation in a direction rather than other ones. A second experiment was performed, introducing knowledge components external to the corpus, i.e. the decomposition in simple glyphs of ligature glyphs, that confirmed the validity of this decomposition and allowed to reinforce the precision of the performed semantic division.

To refine the methodology, a number of variants remain to explore. On the one hand, by changing parameters in the processing chain: changing N, combining 1-grams and 2-grams, exploring other granularity levels for text statistical units such as sentences or cartouches instead of page segments, including n-grams that cross over the borders of 2 successive cartouches, in order to enlighten syntax structure and contexts. On the other hand, by modifying the corpus, i.e. selecting of a homogeneous and continuous sub-corpus as Venus stage tables, or approaching the contexts of a Maya phrase to be deciphered by selecting the text segments containing it as a sub-corpus; or extending the corpus to the codices of Madrid and Paris, and even to inscriptions on ceramics, bringing light on rare glyphs thanks to other contexts. We also plan to test other methodological approaches characterizing texts by "pseudo-n-grams" or *Triggers* [25], relaxing the n-grams strict consecutiveness constraint; or also using glyph itemsets<sup>4</sup> [26], which would relax the sequentiality constraint inside the cartouches, while preserving inter-cartouche sequentiality. These first experiments encourage us to continue further in this direction. Indeed, no breakthrough has emerged from such partial and experimental results, but it appears that a stabilized procedure validated on the whole Maya corpus (the three codices and the ceramics and monument inscriptions) might induce a significant progress in the proportion of well-deciphered glyphs.

This case study illustrates the importance of mining homogeneous semantic contexts out of the data, respectful of both the diverse salience levels, or centralities, of the considered units (features, objects, individuals,...), and tolerant to some degree of polysemy - a central element in a given context may display significant projections in other contexts. Out of linguistic applications, many domains may benefit from this cognition-friendly type of analysis: among other topics, we have applied this methodology to delineating scientific fields and trends, on the basis of co-citation networks [27], as well as co-word analysis, or both [28], by mining large subsets of bibliographic databases.

Another important methodological progress should come when getting rid of the drawbacks of most K-means or probabilistic clustering methods, i.e. initialization-dependance, and the need to determine the "right" number of clusters to be extracted. We have drawn up perspectives in that direction in recent papers of ours: in [29] and [30] we have set up an initialization-independent density clustering method; in [31] and [32] we have described a statistically rigorous randomization test for establishing the intrinsic dimension of a binary dataset, which amounts to the upper bound of the number of clusters in the case of sparse binary data. The present work could then be considered as a feasibility study

<sup>4</sup>A set of two to n glyphs, non necessarily consecutive nor ordered within the same textual unit, for example the cartouche.

in view of an exhaustive processing of the whole Mayan corpus, incorporating our recent advances and resulting in a stable basis for documenting the discussions on controversial or unknown glyphs in the Mayanist community.

#### REFERENCES

- [1] E. Rosh, "Cognitive representations of semantic categories," *Journal of Experimental Psychology*, vol. 104, pp. 192–233, 1975.
- [2] J. C. Bezdek and J. C. Dunn, "Optimal fuzzy partition: A heuristic for estimating the parameters in a mixture of normal distributions," *IEEE Trans. Comput.*, vol. C-24, pp. 835–838, 1975.
- [3] A. Brun, K. Smāli, and J. P. Haton, "Experiment analysis in newspaper topic detection," in *Proceedings of String Processing and Information Retrieval*, La Coruña, Spain, 2000, pp. 55–64.
- [4] E. V. Évréinov, Y. G. Kosarev, and V. A. Oustinov, *Primenenie elektronikh vychislitel'nykh mashin v issledovanii pis'mennosti drevnykh mayia [Utilisation des machines calculer électroniques pour les recherches sur l'écriture des anciens mayas]*. Novossibirsk: Akademia nauk SSSR [Académie des sciences de l'URSS], 1969, 4 vol.
- [5] B. Delprat, "Le codex de dresde: Paléographie et traduction comparée d'un almanach maya du 15<sup>ème</sup> siècle," Ph. D. Dissertation, Institut National des Langues et Civilisations Orientales, Paris, n.p., dissertation in progress.
- [6] B. Delprat and S. Orevkov, "maya – un sistema de composición tipográfica de textos jeroglíficos mayas para la computadora," in *XXI Simposio de investigaciones arqueológicas en Guatemala*, Guatemala de la Asunción, July 2007.
- [7] E. W. Förstemann, *Die Maya-Handschrift der königlichen Bibliothek zu Dresden*. Leipzig: Verlag der A. Naumann'schen Lichtdruckerei, 1880.
- [8] A. Barrera Vásquez and S. Rendón, *El libro de los libros de Chilam Balam*, Mérida, México, 1948, 152 p.
- [9] A. Lelu, M. Hallab, and B. Delprat, "Recherche d'information et cartographie dans des corpus textuels partir des fréquences de n-grammes," in *Actes JADT 1998*, 1998.
- [10] K. Matusita, "Decision rules based on distance for problems of fit, two samples and estimation." *Ann. Math. Stat.*, vol. 26, no. 4, pp. 631–640, 1955.
- [11] B. Escofier, "Analyse factorielle et distances répondant au principe d'équivalence distributionnelle." *Revue de Statistique Appliquée*, vol. 26, no. 4, pp. 29–37, 1978.
- [12] D. Domengès and M. Volle, "Analyse factorielle sphérique: une exploration," *Annales de l'INSEE*, vol. 35, pp. 3–83, 1979.
- [13] C. Rao, "A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance," *Questio*, vol. 19, pp. 23–63, 1995.
- [14] J.-P. Benzécri, *L'analyse des correspondances*. Paris: Dunod, 1973.
- [15] A. Renyi, *Calcul des probabilités*. Paris: Dunod, 1966, 620 p.
- [16] A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using Von Mises-Fisher distributions," *Journal of Machine Learning Research (JMLR)*, vol. 6, pp. 1–39, 2005.
- [17] A. Lelu, "Clusters and factors: Neural algorithms for a novel representation of huge and highly multidimensional data sets," in *New Approaches in Classification and Data Analysis*, E. Diday, Y. Lechevallier, and al., Eds. Springer-Verlag, 1994, pp. 241–248.
- [18] M. Greenacre and T. Hastie, "The geometric interpretation of correspondence analysis," *Journal of the American Statistical Association*, vol. 82, No. 398, pp. 437–447, 1987.
- [19] U. Brandes and S. Cornelsen, "Visual ranking of link structures," *Journal of Graph Algorithms and Applications*, vol. 7:2, pp. 181–201, 2003.
- [20] J. D. Zucker and J. G. Ganascia, "Learning structurally indeterminate clauses," in *Proceedings of the 8th international Conference on ILP*. Springer-Verlag, 1998, pp. 235–244.
- [21] S. G. Zhou and al., "A chinese document categorization system without dictionary support and segmentation processing," *Journal of Computer Research and Development*, vol. 38, no. 7, pp. 839–844, 2001.
- [22] Z. Wei, D. Miao, J. H. Chauchat, and C. Zhong, "Feature selection on chinese text classification using character n-grams," in *RSKT 2008 : the 3rd International Conference on Rough Sets and Knowledge Technology*, may 2008, pp. 500–507.
- [23] Z. Wei, "Avancée en classification multi-labels de textes en langue chinoise," Ph. D. Dissertation, University Lumière2, Lyon, France, may 2010.
- [24] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
- [25] R. Lau, "Maximum likelihood maximum entropy trigger language model," Ph.D. dissertation, Massachusetts Institute of Technology, may 1993.
- [26] M. Cadot and A. Lelu, "Simuler et épurer pour extraire des motifs pertinents," in *QDC2007*, Namur, Belgium, January 2007.
- [27] E. Bassecoulard, A. Lelu, and M. Zitt, "Mapping nanosciences by citation flows: a preliminary analysis," *Scientometrics*, vol. 70, pp. 859–880, 2007.
- [28] M. Zitt, A. Lelu, and E. Bassecoulard, "Hybrid citation-word representations in science mapping: portolan charts of research fields ?" *JASIST*, vol. 62, pp. 19–39, 2010.

- [29] A. Lelu, P. Cuxac, and J. Johansson, "Classification dynamique d'un flux documentaire : une évaluation statique préalable de l'algorithme GERMEN." in *JADT 2006 : 8es Journées internationales d'Analyse statistique des Données Textuelles*, France, April 2006, pp. 617–630.
- [30] P. Cuxac, M. Cadot, and A. Lelu, "Incremental analysis of the evolution of an indexed database: an assessment/correction loop for the choice of algorithms and parameters," *Int. J. of Data Mining, Modelling and Management (IJDMMM)*, 2011, to appear.
- [31] A. Lelu, "Slimming down a high-dimensional binary datatable: relevant eigen-subspace and substantial content," in *Proceedings of COMPSTAT42010 19th International Conference on Computational Statistics - COMPSTAT 2010*, G. S. Yves Lechevallier, Ed. Paris France: Physica-Verlag, 08 2010, pp. 1271–1278.
- [32] A. Lelu and M. Cadot, "Espace intrinsèque d'un graphe et recherche de communautés," in *Actes de la première conférence sur les Modèles et l'Analyse des Réseaux : Approches Mathématiques et Informatique Première conférence sur les Modèles et l'Analyse des Réseaux : Approches Mathématiques et Informatique - MARAMI 2010*, Frédéric Amblard, Ed., Toulouse France, 10 2010, pp. 1–12.

Table II

CLUSTER 1 MAYA TERMS BY DECREASING CENTRALITY. DRESDEN CODEX ENCODED WITHOUT SPLITTING LIGATURES INTO ELEMENTARY GLYPHS

Rank	Maya term	Colonial Yucatec transcription	Meaning and translation	Corresponding folio segments
1	 364/(153.153)	<i>baktun / pictun</i>	20 x20 x18 x20 = 144.000 days cycle, i.e. 400 years	D61AB, D69A
2	 069/(-505.322.-505)  069c/(-505.274.-505)	<i>pawah thul pawah cizin</i>	divinity rabbit divinity death [unsure]	D61AB, D69A
3	 173/112	<i>uinic</i>	man, 20 days cycle	D61AB, D69A
4	 023.153.023)/220	<i>katun</i>	20X18x20=7.200 days cycle, i.e. 20 years	D61AB, D69A
5	 220/009	<i>tun/haab</i>	18x20=360 days cycle, solar year (360+5 jours "unnamed" intercalary days)	D61AB, D69A
6	 105/155.030	<i>pat otoch-ah / pat-ah / kat-ah</i>	place into the house/to form	D52b, D61AB, D69A
7	 054.212	<i>och ixim/ha'</i>	enter [into] maize/water	D31a, D61AB
8	 056/212	<i>ti ixim/ha'</i>	into maize/water	D51a, D52a, D61AB, D69B
9	 060/?705.030	<i>o chicchan-ah</i>	[not understood]	D61AB
10	 204/031	<i>kin</i>	day, sun	D61AB
11	 245.235	<i>yax Ahau</i>	the green lord	D61AB, D69A
12	 804.700	<i>chan ahau</i>	14 Ahau [260 days Tzolkin calendar date]	D31a, D51b, D69A, D70b
13	 411/515	<i>cumku</i>	Maya month Cumku	D31a
14	 075.300/031	<i>yoan / yoan kin</i>	parents of the sun [unsure]	D61AB, D69A

Table III

HIGHEST CENTRALITY MAYA TERMS FOR CLUSTER 5. DRESDEN CODEX WITH LIGATURE GLYPHS SPLITTED INTO DEVELOPED ENCODINGS.

Rank	Maya terms of highest centrality	Colonial Yucatec transcription	Meaning / Theme	Number of bigrams in the cluster
1	 220/009  807.704	<i>tun/haab 8 Kan</i>	year, date in the 260 days <i>tzolkin</i> calendar	19
2	 809.703  911.711	<i>9 Akbal 11 Chuen</i>	dates in the 260 days <i>tzolkin</i> calendar	23
3	 276/010  307/067.504	<i>cimi, kam</i>	death	14
4	 026.172/023  034.319.034/135	<i>u-muc, Ixchel,</i>	its omen, moon divinity	22
5	 (076.234.076)/024  (154.123)/177	<i>zih-an, ahau dzak</i>	born from, accession to power	41