



Codage et classification non supervisée d'un corpus maya : extraire des contextes pour situer l'inconnu par rapport au connu

Mohamed Hallab, Bruno Delprat, Alain Lelu

► To cite this version:

Mohamed Hallab, Bruno Delprat, Alain Lelu. Codage et classification non supervisée d'un corpus maya : extraire des contextes pour situer l'inconnu par rapport au connu. Jean-Marc Petit. Extraction et Gestion de Connaissances - EGC 2010, Jan 2010, Sousse, Tunisie. CEPADUES, E19, pp.573-584, 2010. <hal-00435233v2>

HAL Id: hal-00435233

<https://hal.archives-ouvertes.fr/hal-00435233v2>

Submitted on 23 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.

Codage et classification non supervisée d'un corpus maya : extraire des contextes pour situer l'inconnu par rapport au connu

Mohamed Hallab^{*,**} Bruno Delprat^{***,****} Alain Lelu^{‡,‡‡}

*CNAM, La Défense, 15 Ave d'Alsace, 92400 Courbevoie
mohamed.hallab@yahoo.fr

**Université du 7 Novembre, École Supérieure de Technologie et d'Informatique
4, rue des Entrepreneurs Charguia II - 2035 Tunis-Carthage

***INALCO, École doctorale - 49bis, avenue de la Belle Gabrielle, 75012 Paris
brunodelprat@club-internet.fr

****CÉLIA-CNRS - 7, rue Guy Môquet, 94801 Villejuif cedex

‡Université de Franche-Comté, Laseldi, 25030 Besançon Cedex
alain.lelu@univ-fcomte.fr

‡‡LORIA, Éq. KIWI, Campus Scientifique - BP 239 - 54506 Vandœuvre-lès-Nancy Cedex

Résumé. L'écriture logosyllabique des anciens Mayas comprend plus de 500 signes et est en bonne partie déchiffrée, avec des degrés de certitude divers. Nous avons appliqué au codex de Dresde, l'un des trois seuls manuscrits qui nous soient parvenus, codé sous L^AT_EX avec le système mayaT_EX, notre méthode de représentation graduée, par apprentissage non supervisé hybride entre clustering et analyse factorielle oblique, sous la métrique de Hellinger, afin d'obtenir une image nuancée des thèmes traités : les individus statistiques sont les 212 segments de folio du codex, et leurs attributs sont les 1687 bigrammes de signes extraits. Pour comparaison, nous avons introduit dans cette approche endogène un élément exogène, la décomposition en éléments des signes composites, pour préciser plus finement les contenus. La rétro-visualisation dans le texte original des résultats et expressions dégagées éclaire la signification de certains glyphes peu compris, en les situant dans des contextes clairement interprétables.

1 Introduction et problématique




L'écriture logosyllabique des anciens Mayas, en usage pendant plus de 13 siècles, nous est parvenue au travers de riches inscriptions sur des monuments, des céramiques et trois almanachs divinatoires, qui constituent néanmoins un volume faible de textes disponibles : trois manuscrits et quelques milliers d'inscriptions courtes découvertes.

L'objectif du travail présenté ici est de dégager les principaux contextes sémantiques d'usage des glyphes, dans l'esprit de la sémantique des prototypes (Rosh, 1975), de façon à mettre en contexte commun des glyphes élucidés et ceux qui le sont moins ou pas du tout. A terme, il pourrait déboucher sur la mise à disposition de la communauté scientifique mayaniste de


cette mise en contexte pour l'ensemble du corpus maya disponible. Nous avons choisi pour ce premier corpus codé informatiquement une méthode de représentation : 1) non supervisée, afin que des co-occurrences de glyphes, ou autres attributs des textes, naissent des contextes interprétables à l'aune de ce qu'on connaît déjà, 2) floue : chaque unité statistique, texte élémentaire ou attribut, se voit attribuer une valeur de centralité plus ou moins forte dans les divers contextes dégagés. Ainsi, certains éléments mal représentés dans l'analyse auront partout des valeurs faibles ; d'autres, polysémiques ou syntaxiques, pourront être centraux dans plusieurs contextes, c'est-à-dire avoir une valeur forte dans plusieurs classes ; d'autres encore à signification univoque seront centraux dans une classe seulement, ce qui démarque notre démarche du «fuzzy clustering» (Bezdek et Dunn, 1975) car la somme de nos centralités d'une unité statistique dans les diverses classes n'est pas contrainte à être égale à 1, une centralité ne traduit pas une incertitude sur l'appartenance à une classe (une probabilité), mais une participation à la construction d'un contexte, 3) compatible avec la rareté relative des sources disponibles, par contraste avec l'approche des modèles statistiques de langage qui demandent de collationner des millions d'occurrences (Brun et al., 2000).

2 Principe de l'écriture et de son codage sous L^AT_EX

Principes généraux de l'écriture maya Le signe d'écriture élémentaire est le glyphe. Un ou plusieurs glyphes sont assemblés ensemble pour remplir harmonieusement l'espace rectangulaire prédéfini d'un cartouche. Les textes des manuscrits mayas sont organisés en blocs de 2 à 12 cartouches qui constituent autant de phrases. Selon le nombre de cartouches dont il disposait pour écrire une phrase plus ou moins longue dans la page d'almanach, le scribe pouvait entasser ou étaler les éléments dans les cartouches pour ne pas laisser de case vide et obtenir une belle mise en page.









Par l'analyse des corpus des codex de Dresde et Madrid, nous avons identifié la constitution des cartouches glyphiques complets obtenus par la composition de 1 à 5 éléments de base. Les affixes, comme 031  *ni*, s'organisent en tournant et par symétries, selon une règle déterminée, autour des éléments centraux 204  dont l'orientation est fixe. Les affixes sont le plus souvent des signes à valeur syllabique que l'on peut combiner ensemble ou avec un élément central. Les éléments centraux sont le plus souvent des logogrammes correspondant à un morphème ou mot, dont la lecture est globale comme 204  *KIN* soleil, jour. La règle générale d'orientation des affixes est la suivante :




(063) *te*     Par exemple :    

Un cartouche glyphique complet correspond souvent à une entrée lexicale avec les affixes grammaticaux qui la précèdent et la suivent 204.031  *KIN-ni* soleil, jour, mais il peut aussi parfois correspondre à deux mots s'ils sont courts, ou encore plus rarement à une partie d'un terme qui s'écrit sur deux cartouches.

Saisie et composition informatique des glyphes mayas sous mayaT_EX Dans les années 1960 des cryptologues et historiens soviétiques (Évréïnov et al., 1969) ont entrepris un codage

informatique de la version cursive de cette écriture qui comprend environ 500 glyphes élémentaires, et son exploitation permise par l'état de l'art de l'époque, travail tombé dans l'oubli depuis.

Préalable au codage informatique de textes mayas est l'analyse structurelle de la forme écrite des Codex, qui comprend l'énonciation de règles de composition des signes hiéroglyphiques de base dans le cartouche maya, la définition d'une grammaire graphique, réalisés dans le cadre d'une thèse de Doctorat en cours (Delprat, np). Un outil informatique original développé sous T_EX de saisie et édition de textes hiéroglyphiques mayas mayaT_EX (Delprat et Orevkov, 2007) est utilisé pour la composition de la paléographie et des lexiques intégrés à la thèse. Les deux principaux opérateurs de composition des glyphes au sein du cartouche maya sont le point "." qui associe deux éléments 117  et 260  en les juxtaposant dans le sens gauche-droite 117.260 , et la barre oblique "/" qui place un élément 400  au dessus de l'autre 010  pour donner 400/010.030 . L'opérateur "." ou "/" agit sur le sous-ensemble de glyphes mis entre parenthèses (154.123) , pour donner (154.123)/177 .

Les ligatures Un ou plusieurs éléments affixe ou central peuvent s'inscrire à l'intérieur d'un élément central géométrique ou, le plus souvent, d'une tête au lieu de lui être simplement accolé ; il y a alors ligature en un seul dessin. La ligature ainsi formée est matérialisée dans le catalogue par un élément graphique spécifique avec numéro propre 373  *cacau* D7c(2), qui en fait se décompose : 369<023/023>. L'opérateur < > indique que les deux affixes 023  sont inscrits au centre de 369 . Dans les textes, les deux formes : liée (décrite par 1 code), et non liée (décrite en 2 à 3 codes) sont équivalentes et constituent des variantes graphiques.

3 Le codex de Dresde et son déchiffrement

Notre corpus Le codex de Dresde, l'un des trois seuls manuscrits mayas qui nous soient parvenus et datant probablement du 15^e siècle, est constitué d'un ensemble de 76 almanachs de 5 types principaux : almanachs divinatoires du calendrier *tzolkin* de 260 jours consacrés à diverses divinités, prophéties de l'année solaire *haab* de 360 jours plus 5 jours intercalaires et des *katuns* ou cycles de 52 ans, almanachs des quatre directions cardinales consacrés à Chac, le dieu de l'eau, tables astronomiques telles les phases de Vénus et les éclipses de soleil et de lune, et almanachs des cérémonies de la nouvelle année et du déluge associé au cycle *katun* de 52 ans. Ces textes sont en général indépendants les uns des autres et non les chapitres successifs d'un livre occidental à lire d'un début jusqu'à une fin.

Les 74 folios, numérotés par (Förstemann, 1880) et codés selon les numéros du catalogue de (Évréinov et al., 1969) complété pour mayaT_EX, ne se lisent pas linéairement l'un après l'autre. En effet, chaque almanach divinatoire ou texte avec table astronomique du codex est peint transversalement sur plusieurs folios, par exemple les parties supérieures folios 4a, 5a, 6a... à 10a. Les unités statistiques de texte considérées ici sont les 212 segments de folio tels que définis par les scribes mayas en général séparés par une ligne rouge, qui correspondent chacun

en règle générale à une partie supérieure, centrale ou inférieure de folio. Chacun comprend 1 à 7 blocs de cartouches hiéroglyphiques, formant autant de phrases. Du point de vue quantitatif, le corpus comporte 9938 occurrences de glyphes, dont 411 distincts¹.

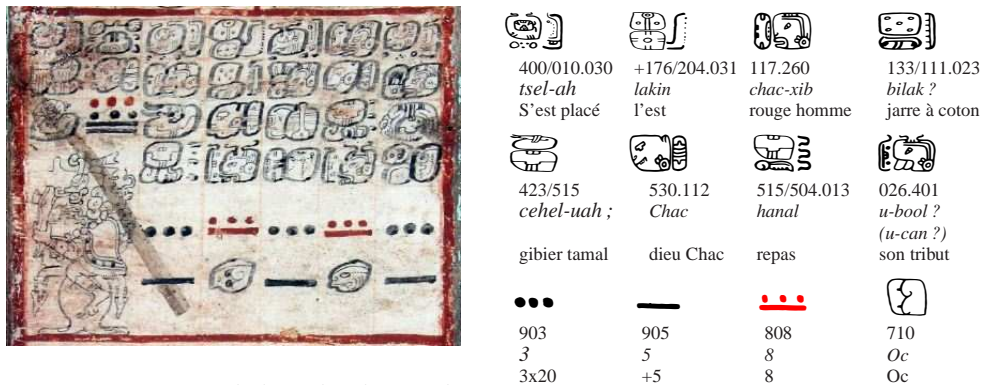


FIG. 1 – Segment 30b du codex de Dresde







La Figure 1 donne la reproduction du segment médian numéroté Page 30b avec la paléographie hiéroglyphique correspondante au 2^e bloc de texte depuis la gauche codé sous mayaTeX. Cet almanach fait partie d'une série consacrée aux occupations de Chac, le dieu de la pluie et de l'eau qui prend la forme des avatars des quatre orient. Les offrandes correspondantes sont un plat de viande et un tribut d'une couleur particulière pour chacune des directions :

À l'est s'est placé Chac en homme rouge ; son tribut est un repas de tamales de gibier et une jarre de coton. 65 [jours jusqu'au] 8 Oc.

Son déchiffrement Il bénéficie de facteurs favorables : les langues mayas sont encore parlées de nos jours, et l'on dispose des prophéties des *Chilam Balams* (Barrera Vásquez et Rendón, 1948) qui sont partiellement la retranscription en écriture latine yucatèque de textes divinatoires semblables aux trois codex divinatoires hiéroglyphiques. La signification des signes logo-syllabiques, appelés glyphes, est établie de façon certaine pour plus d'un cinquième, et plausible pour une bonne moitié.

Ils se répartissent en trois principaux types. Le logogramme, signe morphémique aux valeurs à la fois sémantiques et phonétiques, le plus souvent monosyllabiques de forme consonne-voyelle-consonne (CVC), comme le glyphe 204 représentant une fleur à quatre pétales, symbole du soleil et de valeur phonétique KIN soleil, jour. Le syllabogramme, notant une syllabe de type CV ou VC, comme le glyphe +176 de valeur phonétique la est qui entre dans l'écriture de différents mots avec un autre syllabogramme ou un logogramme comme +176/204 la-KIN (l'est). Les syllabogrammes sont dérivés de logogrammes de valeur CV(C) dont la deuxième consonne est faible. Le complément phonétique, signe phonétique





1. Ex. : il y 39 glyphes qui apparaissent 2 fois.

indiquant la valeur CV, VC ou une partie de la valeur C(V) du logogramme auquel il est associé, comme pour le mot 204.031  où 031  est le complément phonétique ni ou in est postfixé à KIN (soleil, jour) dont il vient confirmer la valeur de lecture. Un mot peut être écrit par différentes combinaisons de glyphes notant la même valeur phonétique, comme pour lakin (l'est) indifféremment écrit par les allographes +176/204 , 176/204.031 , 204/176.031  ou 031.+176/204 .

4 La chaîne de traitements

Le corpus codé sous *mayaTeX* a d'abord fait l'objet d'un pré-traitement pour extraire les n-grammes de glyphes élémentaires. Ensuite, un algorithme de clustering décrit plus loin et programmé sous *Scilab*, fournit pour chaque classe les listes ordonnées et valuées des n-grammes et documents caractéristiques de cette classe. Ces listes sont retraitées pour aligner les cartouches mayas du texte d'origine avec les n-grammes, et enfin les résultats sont visualisés en écriture maya à l'aide de *mayaTeX* pour permettre leur interprétation linguistique.


4.1 Pré-traitement

Choix de découpage des unités textuelles Le choix du découpage en unités statistiques de texte – unités dont la comparaison est la raison d'être de l'analyse - est un choix de granularité de l'analyse : trop fin, par exemple ici au niveau du cartouche ou du glyphe, il privilégie les éléments syntaxiques, ce qui n'est pas notre préoccupation présente ; plus grossier, il privilégie les éléments sémantiques, mais le risque est d'avoir trop peu d'éléments pour qu'une classification au niveau de finesse souhaité en ressorte. Le compromis fait ici est de prendre en considération les segments de pages définies par les scribes mayas, sachant que les textes des segments peuvent se poursuivre sur plusieurs pages. Le nombre de segments de pages est compatible a priori avec une granularité d'analyse d'une dizaine de classes sémantiques, permettant d'aller au-delà des divisions triviales et connues du texte (phases de Vénus, prévision des éclipses, etc.). Une division plus fine, au niveau des phrases mayas a également été expérimentée, pour valider notre parti-pris de découpage, et prolonger l'analyse sur le plan syntaxique. Extrait du segment de texte 5b :  ...  => phrase  => cartouche  => glyphe

Pourquoi des n-grammes ? Choix de n Pour les expériences présentées ici nous avons pris le parti de caractériser chaque portion de texte par un vecteur de fréquences de bigrammes de glyphes, tels qu'ils ressortent de leur codage par *mayaTeX* : la combinatoire observée de ces bigrammes, largement inférieure à 500^2 , est maîtrisable dans l'état de l'art informatique actuel sans avoir à faire appel à la compression du nombre de codes par H-coding, comme nous avons pu le faire dans d'autres contextes d'application par le passé (Lelu et al., 1998) ; notre choix a été de ne pas faire franchir aux bigrammes les frontières des cartouches, qui constituent le plus souvent des expressions séparées. Les bigrammes constituent une façon souple et minimale de traduire la séquentialité du texte et correspondent le plus souvent à une partie de cartouche, de 3 à 5 signes, et donc de mot ou expression maya qui sont donnés après chaque bi-gramme dans les tableaux.

Notre logiciel de présentation des résultats, par post-traitement des sorties (codées) de l'étape de clustering, a dû être adapté aux spécificités de l'écriture maya. Une des difficultés que nous avons dû résoudre est la suivante : les codes des glyphes sont entourés par des codes de positionnement (au-dessus/au-dessous, type de symétrie utilisée par rapport au glyphe de référence, ...) qu'il est facile de filtrer pour obtenir des bi-grammes de glyphes sans cette information. En sens inverse, il est indispensable de rétablir ces éléments pour visualiser chaque bigramme important d'une classe au sein de son contexte graphique d'origine : nous avons choisi de présenter à l'utilisateur mayaniste, pour chaque bi-gramme important d'un cluster, tous les cartouches différents dans lesquels celui-ci intervient, réalisant en quelque sorte une concordance au niveau de chaque classe.

Extraction des Ngrammes Les textes d'origine sont transformés par réduction des caractères de positionnement / :< et filtration des attributs d'orientation (*?~!+@|>. Une fenêtre de longueur 7 (2 glyphes mayas séparés par un point) se déplace le long du texte, 4 caractères à la fois. L'espace arrête le balayage courant par la fenêtre de N=2 glyphes, puis le réinitialise.

<u>Exemple</u>	Texte paléographié :	
	Texte codé :	990.172/056 (154.123)/306 *002c
	Texte transformé :	990.172.056 154.123.306 002c
	Bigrammes extraits :	990.172 , 172.056 , 154.123 , 123.306

4.2 Processus d'apprentissage non supervisé

Distance et cosinus distributionnels Une lignée ancienne de travaux (Matusita, 1955) (Escofier, 1978) (Domengès et Volle, 1979) (Rao, 1995) s'est intéressée à ce que certains auteurs appellent distance distributionnelle et d'autres distance de Hellinger) : il s'agit de la distance euclidienne, classique (équipondération des dimensions), entre les 2 points t_1 et t_2 , de coordonnées fournies par les vecteurs \mathbf{z}_{t_1} et \mathbf{z}_{t_2} situés sur l'hypersphère unité dans l'espace des I mots, et représentant chacun une unité de découpage textuel, définis par la transformation suivante sur les données : $\mathbf{z}_{t_1} : \left\{ \sqrt{\frac{x_{it_1}}{x_{.t_1}}} \right\}$; $\mathbf{z}_{t_2} : \left\{ \sqrt{\frac{x_{it_2}}{x_{.t_2}}} \right\}$ où x_{it} désigne la fréquence du mot i dans le document t , et $x_{.t}$ le nombre total de mots du document t . La distance distributionnelle $Dd(t_1, t_2)$ entre les textes t_1 et t_2 est donc :

$$Dd(t_1, t_2) = \|\mathbf{z}_{t_1} - \mathbf{z}_{t_2}\| \quad \text{où } \|\mathbf{x}\| \text{ désigne la norme euclidienne du vecteur } \mathbf{x}.$$

Cette distance est la longueur de la corde correspondant à l'angle $(\mathbf{z}_{t_1}, \mathbf{z}_{t_2})$ - égale au plus à 2 quand ces 2 vecteurs sont opposés, égale à $\sqrt{2}$ quand ils sont orthogonaux. Cette distance semble triviale et arbitraire en apparence (pourquoi cette normalisation insolite plutôt que la normalisation classique $\left\{ \frac{x_{it}}{\|\mathbf{x}_t\|} \right\}$), mais elle jouit de propriétés intéressantes : 1) Contrairement à la distance du khi-deux utilisée en Analyse Factorielle des Correspondances (Benzécri, 1973), ou AFC, elle peut prendre en compte des vecteurs ayant des composantes négatives, propriété utile pour certains types de codage «symétriques» (comme Oui, Non, Ne sait pas) ou pour des tableaux de flux orientés - économiques, physiques, ... 2) Elle est liée à la mesure du gain d'information de Renyi d'ordre $\frac{1}{2}$ (Renyi, 1966) apporté par une distribution \mathbf{x}_q quand on connaît la distribution \mathbf{x}_p : $I^{(1/2)}(\mathbf{x}_q/\mathbf{x}_p) = -2 \log_2(\cos(\mathbf{z}_p, \mathbf{z}_q)) = -2 \log_2\left(1 - \frac{Dd^2}{2}\right)$ 3) Rapide à calculer dans le cas des données textuelles, où les vecteurs \mathbf{z}_t sont très creux, 4)

et surtout, (Escofier, 1978) et (Domengès et Volle, 1979) ont montré qu'elle satisfaisait à la même propriété d'équivalence distributionnelle que la distance du χ^2 utilisée en AFC : si on fusionne deux descripteurs de mêmes profils relatifs, les distances entre les unités textuelles sont inchangées. En d'autres termes, dans le cas où les descripteurs sont des mots et les unités décrites des textes, cette propriété assure la stabilité du système des distances entre textes au regard de l'éclatement ou du regroupement de mots de distributions proches.

Notre méthode de classification non supervisée Les principes à l'oeuvre dans notre méthode de clustering (Lelu, 1994) sont 1) de transformer le nuage de données brut des documents en un nuage de données normalisé à la surface de l'hypersphère unité au moyen de la transformation : $x_{ij} \rightarrow \sqrt{\frac{x_{ij}}{x_{i.}}}$ où $x_{i.}$ est la somme du vecteur-document \mathbf{x}_i , 2) d'éclater ce nuage en K sous-nuages, chacun muni d'un axe issu du centre de l'hypersphère. Chaque point appartient alors au cluster sur l'axe duquel sa projection est maximale.

L'axe de chaque sous-nuage est défini comme le premier vecteur propre extrait par Analyse Factorielle Sphérique (AFS) (Domengès et Volle, 1979) (option «différence au tableau nul») : en notation matricielle, si \mathbf{X} est le tableau de données (documents x attributs), dont la somme en ligne est $x_{i.}$ et en colonne $x_{.j}$, si $\mathbf{D}_r^{-\frac{1}{2}}$ est la matrice diagonale de $\{x_{i.}^{-\frac{1}{2}}\}$ (respectivement $\mathbf{D}_c^{-\frac{1}{2}}$ avec $\{x_{.j}^{-\frac{1}{2}}\}$), la décomposition aux valeurs singulières (SVD) de $\mathbf{X}^{\frac{1}{2}} = \{x_{ij}^{\frac{1}{2}}\}$ s'écrit : $\mathbf{X}^{\frac{1}{2}} = \mathbf{U}\mathbf{D}\mathbf{V}'$ où le signe prime indique la transposée d'une matrice, et où les conditions $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}$ sont vérifiées. Les facteurs s'écrivent : $\mathbf{F} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}\mathbf{D}$ et $\mathbf{G} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}\mathbf{D}$

À noter que ce processus est formellement lié à l'analyse des correspondances (AFC), où on applique la SVD à la matrice transformée $\mathbf{Q} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{X}\mathbf{D}_c^{-\frac{1}{2}}$ ce qui entraîne : $\mathbf{Q} = \mathbf{U}_{ca}\mathbf{D}_{ca}\mathbf{V}'_{ca}$, les facteurs AFC s'écrivant : $\mathbf{F}_{ca} = x_{.j}^{\frac{1}{2}}\mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}_{ca}\mathbf{D}_{ca}$ et $\mathbf{G}_{ca} = x_{i.}^{\frac{1}{2}}\mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}_{ca}\mathbf{D}_{ca}$

En termes géométriques, l'AFC projette le nuage brut des données sur la surface d'un «simplexe étiré» (Greenacre et Hastie, 1987) dont le barycentre est pointé par le premier facteur, un vecteur trivial de uns (la première valeur propre, associée, est égale à un). Ce qui contraste avec l'AFS, où les facteurs-documents, c-à-d les projections des documents sur le premier axe, définissent les indices de centralité de ces documents², et le carré de la première valeur propre λ_1^2 définit la portion de la somme $x_{.}$ du tableau de données due à la reconstitution de premier ordre de $\mathbf{X} \simeq \left\{x_{ij} = \frac{1}{\lambda_1}\mathbf{F}_1^2(i)\mathbf{G}_1^2(j)x_{i.}x_{.j}\right\}$, où $\mathbf{F}_1(i)$ représente la i -ème composante du premier facteur-ligne, et de façon symétrique pour $\mathbf{G}_1(j)$.

Dans l'application présentée ci-dessous, la somme pour tous les clusters du carré de leur première valeur propre rend compte de 25.21 % des données. Cette approche permet, à partir d'un clustering strict des unités textuelles, d'obtenir des représentations nuancées : typicité (*typicality*) d'un segment textuel au sein de plusieurs contextes sémantiques, et non d'un seul, spécificité (*cue-validity*) de chaque bigramme dans les divers contextes, et relations de dualité entre ces indices, inexistantes à notre connaissance dans les autres méthodes de clustering strict ou flou.

2. Il découle directement des propriétés des vecteurs propres que le premier vecteur propre de la table des cosinus $\mathbf{X}^{\frac{1}{2}}\mathbf{D}_c^{-1}\mathbf{X}^{\frac{1}{2}'}$ est \mathbf{F}_1 . Ce vecteur peut être interprété comme l'ensemble des «centralités spectrales» (Brandes, 2003) des noeuds du graphe valué dont la matrice d'adjacence est cette table des cosinus.


4.3 Post-traitement : présentation des résultats




Comme on le détaillera plus bas, notre algorithme de classification non supervisée fournit pour chaque classe les listes ordonnées et valuées des bigrammes et documents caractéristiques de cette classe. Pour pouvoir interpréter le thème traité dans une classe, nous devons disposer, en plus des intitulés de ses documents, de la liste de ses mots (cartouches dans notre cas) les plus marquants. La construction de cette liste pour une classe donnée nécessite d'effectuer une 2^e passe sur les documents de cette classe. De ce fait, nous extrayons à la fois le bigramme et le cartouche correspondant (expression dans le texte d'origine délimitée par 2 espaces et dont le numéro de séquence dans le texte d'origine correspond à celui du texte réduit).





Les tableaux des classes obtenues sont présentés par ordre de centralité décroissante des expressions mayas pour chaque classe. Nous avons regroupé ensemble les bi-grammes correspondants aux mêmes expressions.

5 Expériences et résultats

Plusieurs séries d'expériences ont été réalisées, et nous nous limitons ici à en présenter deux, réalisées avec le même paramétrage de pré-traitement et de classification non supervisée, mais avec un codage du corpus différent : sans décomposer les glyphes-ligatures complexes, puis en décomposant ceux-ci en glyphes élémentaires simples. La table de décomposition des ligatures constitue l'information exogène introduite dans la deuxième expérience.

Suite à quelques essais, le nombre de classes demandées (10) est apparu comme un compromis raisonnable entre le nombre de sections analysées (212) et la finesse d'analyse attendue. La recherche des bi-grammes est faite parmi les glyphes élémentaires à l'intérieur de chaque cartouche maya, sans lien avec les cartouches voisins. De ce fait, les cartouches à un seul élément sont ignorés. De plus, notre expérience visant l'analyse des parties textuelles du codex et pas celle des tables de calculs astronomiques ou calendaires, n'ont pas été pris en compte les cartouches mayas ne contenant que des nombres (en base 20) rouges de la série 8xx  .


...  ou noirs de la série 9xx  ... Par exemple, le couple 808/917  sera ignoré.


Par contre, les combinaisons d'un nombre (séries 8xx ou 9xx) avec un glyphe autre qu'un nombre sont prises en compte, y compris les dates du calendrier *tzolkin* de 260 jours qui comprennent un chiffre rouge de 1 à 13 avec un des 20 jours de la série 7xx    

...    , comme par exemple 807.704  ou 908.255/220 .

5.1 Sans information exogène




Les ligatures de glyphes élémentaires n'ont pas été décomposées en leurs éléments constitutifs et apparaissent comme des glyphes à part entière. La Table 3 en Annexe présente une classe de 46 bi-grammes d'où se dégagent 21 expressions ou mots hiéroglyphiques, colorées en rouge au sein des cartouches dans la table des expressions mayas et bi-grammes de la classe. En voici par rang de centralité, en l'état actuel du déchiffrement de l'écriture maya, la translittération, traduction et segments correspondants du corpus. Sont explicitées seules les 14 expressions renvoyant à plusieurs occurrences dans le corpus.











Nous constatons que sont associés au premier plan dans cette classe les termes mayas des cycles du compte long maya avec *kin* 1 jour, *uinic* 20 jours, *tun/haab* 360 jours ou 1 an, *katun* 20 ans, *baktun/pictun* 400 ans. En deuxième plan ressort la date origine 4 *Ahau* 8 *Cumku* 

 du compte long maya. Différentes parties de l'almanach du Nombre serpent en compte long et table des multiples de 91 jours constituent l'essentiel de cette classe (6 documents sur les 10 de la classe), mais il y manque environ la moitié des sections de cet almanach. Les sections 61AB et 69A de même structure et à textes très semblables et parallèles avec dates en compte long ont bien été étroitement associées (almanach du Nombre serpent en compte long et table des multiples de 91 jours). Une centralité forte est donnée aux bi-grammes des sections 61AB et 69A présentant une seule occurrence dans le corpus au détriment des autres documents de la classe. La section 31b de l'almanach des dates mythiques et historiques est une reprise résumée de l'almanach précédent, et son inclusion dans la classe 1 est pertinent.






D'autre part, nous observons qu'il s'effectue un tri thématique des segments de textes dans les classes qui coïncide bien avec leur répartition dans les différents almanachs du codex.

5.2 Avec information exogène : le codage développé des glyphes-ligatures

Un nombre important des formes glyphiques (107 sur les 402 signes hormis les chiffres de la police maya employée) sont des ligatures, et à leur saisie globale en un code leur lien avec leurs éléments constitutifs est perdu pour notre méthode de classification supervisée, alors que pourtant visuellement il apparaît d'emblée. Par exemple : en codage global 455  et codage développé 454/111 , qui correspond en fait à la fusion des deux glyphes superposés  de bi-gramme associé 454.111.

Classe	Expressions de centralité forte	Yucatèque romanisé	Signification / Thématique	Bigrammes de la classe	
1	 220/009	 807.704	<i>tun/haab</i> 8 <i>Kan</i>	année, date calendrier de 260 jours <i>tzolkin</i>	19
2	 809.703	 911.711	9 <i>Akbal</i> 11 <i>Chuen</i>	dates du calendrier de 260 jours <i>tzolkin</i>	23
3	 276/010	 307/067.504	<i>cimi, kam</i>	la mort	14
4	 026.172/023	 034.321.034/135	<i>u-muc,</i> <i>Ixchel,</i>	son présage, déesse de la lune	22
5	 (076.234.076)/024154.123)/177		<i>zih-an,</i> <i>ahau dzak</i>	né de, accession au pouvoir	41

TAB. 2: Expressions maya de la Classe 1 par centralité. Codex de Dresde avec codage développé des glyphes-ligatures.

Comme pour l'expérience précédente, mais de façon plus tranchée, les 10 classes demandées font ressortir avec centralité forte les termes recouvrant des notions connexes, comme l'illustre la Table 2 pour les Classes 1 à 5. La formation de classes de contenu selon les expressions mayas est précisée par la décomposition des ligatures, qui enrichit le nombre de contextes et occurrences de glyphes élémentaires sinon peu fréquents dans le corpus (typiquement avec moins de 10 occurrences). Egalement, cela permet de mieux discriminer dans le contexte du corpus les affixes grammaticaux des glyphes signifiants, lorsqu'un affixe comme 030  -ah la marque de l'accompli, est écrit fusionné avec le verbe sur lequel il porte en une ligature 374/318.030  ou 318.030 , au lieu d'être juxtaposé séparément en  ou .

6 Conclusion et perspectives

Le codage du Codex de Dresde sous maya \TeX a permis d'initier une série d'expériences de mise en évidence de classes sémantiques par apprentissage non supervisé. Un clustering à résultats nuancés, grâce à notre méthode originale, a été effectué sur la base d'un découpage du texte en portions de pages, caractérisées par leur profil de bigrammes maya. Cette analyse a confirmé le regroupement des glyphes bien élucidés en classes sémantiques déjà connues, et pour d'autres plus sujets à controverse, elle a fait pencher l'interprétation dans une direction plutôt que d'autres. Une deuxième expérience faite en introduisant une connaissance externe au corpus, à savoir la décomposition en glyphes simples des glyphes-ligatures, a confirmé la validité de cette décomposition et a permis de renforcer la précision du découpage sémantique effectué.

Pour affiner la méthode, bon nombre de variantes restent à explorer. D'une part, en modifiant le paramétrage de la chaîne de traitement : changement de N, combiner les 1-grammes et les 2-grammes, exploration d'autres niveaux de granularité des unités statistiques de textes telles les phrases ou cartouches au lieu des segments, inclure les N-grammes à cheval entre 2 cartouches successifs pour éclairer la syntaxe. D'autre part, en agissant sur le corpus : choix d'un sous-corpus homogène et continu comme les tables des phases de Vénus, approche des contextes d'un terme maya à déchiffrer en sélectionnant comme sous-corpus les segments de texte le contenant, étendre le corpus aux codex de Madrid et Paris, voir à des inscriptions sur céramiques pour éclairer par d'autres contextes des glyphes rares. Nous envisageons aussi de tester d'autres approches méthodologiques en décrivant les textes par des «pseudo-N-grammes» ou *triggers* (Lau, 1993), relâchant la contrainte de stricte consécuité des N-grammes ; ou encore par des motifs de glyphes³ (Cadot et Lelu, 2007), qui relâcheraient la contrainte de séquentialité au sein des cartouches, tout en préservant une séquentialité inter-cartouches.

Nos premières expériences nous encouragent à poursuivre dans cette voie.

Références

Barrera Vásquez, A. et S. Rendón (1948). *El libro de los libros de Chilam Balam*. Mérida, México. 152 p.

3. Ensemble de deux à n glyphes non nécessairement consécutifs ni ordonnés au sein de la même unité textuelle, par exemple le cartouche.

- Benzécri, J.-P. (1973). *L'analyse des correspondances*. Paris : Dunod.
- Bezdek, J. C. et J. C. Dunn (1975). Optimal fuzzy partition: A heuristic for estimating the parameters in a mixture of normal distributions. *IEEE Trans. Comput. C-24*, 835–838.
- Brandes, U. e. C. S. (2003). Visual ranking of link structures. *Journal of Graph Algorithms and Applications* 7:2, 181–201.
- Brun, A., K. Smaïli, et H. J.P. (2000). Experiment analysis in newspaper topic detection. In *Proceedings of String Processing and Information Retrieval*, La Coruña, Spain, pp. 55–64.
- Cadot, M. et A. Lelu (2007). Simuler et épurer pour extraire des motifs pertinents. In *QDC2007 Namur*, La Coruña, Spain.
- Delprat, B. (n.p.). *Le codex de Dresde : Paléographie et traduction comparée d'un almanach maya du 15^esiècle*. Thèse de doctorat, Institut National des Langues et Civilisations Orientales, Paris. (thèse en cours).
- Delprat, B. et S. Orevkov (23-27 juillet 2007). maya \TeX – un sistema de composición tipográfica de textos jeroglíficos mayas para la computadora. In *XXI Simposio de investigaciones arqueológicas en Guatemala*, Guatemala de la Asunción.
- Domengès, D. et M. Volle (1979). Analyse factorielle sphérique : une exploration. *Annales de l'INSEE* 35, 3–83.
- Escofier, B. (1978). Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. *Revue de Statistique Appliquée* 26(4), 29–37.
- Förstemann, E. W. (1880). *Die Maya-Handschrift der königlichen Bibliothek zu Dresden*. Leipzig : Verlag der A. Naumann'schen Lichtdruckerei.
- Greenacre, M. et T. Hastie (1987). The geometric interpretation of correspondence analysis. *Journal of the American Statistical Association* 82, No. 398, 437–447.
- Lau, R. (1993). *Maximum Likelihood Maximum Entropy Trigger Language Model*. Ph. D. thesis, Massachusetts Institute of Technology.
- Lelu, A. (1994). *New Approaches in Classification and Data Analysis*, Chapter Clusters and Factors: Neural Algorithms for a Novel Representation of Huge and Highly Multidimensional Data Sets, pp. 241–248. Berlin: Springer-Verlag.
- Lelu, A., M. Hallab, et B. Delprat (1998). Recherche d'information et cartographie dans des corpus textuels à partir des fréquences de n-grammes. In *Actes JADT 1998*.
- Matusita, K. (1955). Decision rules based on distance for problems of fit, two samples and estimation. *Ann. Math. Stat.* 26(4), 631–640.
- Rao, C. (1995). A review of canonical coordinates and an alternative to correspondence analysis using hellinger distance. *Questio* 19, 23–63.
- Rényi, A. (1966). *Calcul des probabilités*. Paris : Dunod. 620 p.
- Rosh, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology* 104, 192–233.
- Évréïnov, E. V., Y. G. Kosarev, et O. V. A. (1969). *Primenenie elektronikh vychislitel'nykh mashin v issledovanii pis'mennosti drevnykh mayia [Utilisation des machines à calculer électroniques pour les recherches sur l'écriture des anciens mayas]*. Novossibirsk : Akademia nauk SSSR [Académie des sciences de l'URSS]. 4 vol.

Annexe : un exemple de classe de bi-grammes mayas


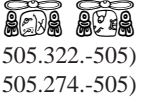










Rang	Expression	Translittération en yucatèque colonial	Signification et traduction	Segments correspondants
1	 364/(153.153)	<i>baktun / pictun</i>	cycle de 20 x20 x18 x20 = 144.000 jours ou 400 ans	D61AB, D69A
2	 069/(-505.322.-505) 069c/(-505.274.-505)	<i>pawah thul pawah cizin</i>	divinité lapin divinité mort [incertain]	D61AB, D69A
3	 173/112	<i>uinic</i>	homme, cycle de 20 jours	D61AB, D69A
4	 023.153.023)/220	<i>katun</i>	cycle de 20X18x20=7.200 jours ou 20 ans	D61AB, D69A
5	 220/009	<i>tun/haab</i>	cycle de 18x20=360 jours, année solaire (360+5 jours intercalaires "sans nom")	D61AB, D69A
6	 105/155.030	<i>pat otoch-ah / pat-ah / kat-ah</i>	mis en la maison/former	D52b, D61AB, D69A
7	 054.212	<i>och ixim/ha'</i>	entrer [dans le] maïs/l'eau	D31a, D61AB
8	 056/212	<i>ti ixim/ha'</i>	dans le maïs/l'eau	D51a, D52a, D61AB, D69B
9	 060/?705.030	<i>o chicchan-ah</i>	[non compris]	D61AB
10	 204/031	<i>kin</i>	jour	D61AB
11	 245.235	<i>yax Ahau</i>	le Seigneur vert	D61AB, D69A
12	 804.700	<i>chan ahau</i>	le 4 Ahau [date du calendrier Tzolkin de 260 jours]	D31a, D51b, D69A, D70b

TABLE 3: Expressions maya de la Classe 1 par centralité. Codex de Dresde codé sans décomposition des ligatures en glyphes élémentaires.

Summary

The ancient Maya logo-syllabic writing comprises more than 500 signs 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 between clustering and oblique factorial analysis, following Hellinger metrics, in order to obtain a nuanced image of themes dealt with : statistical entities are the 212 codex folio segments, and their attributes are the 1687 extracted bi-grammes of signs. For comparison, we introduced to this endogenous approach an exogenous element, i.e. the splitting into elements of composed signs, for a finer explication of the contents. The retro-visualisation of results and outlined expressions into the original text bring new light on the meaning of some little understood glyphs, placing them in clearly understandable contexts.