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

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
[***]Institut National des Langues et Civilisations Orientales, École doctorale
49bis, avenue de la Belle Gabrielle, 75012 Paris
brunodelprat@club-internet.fr
[****]Centre d'Étude des Langues Indigènes d'Amérique-CNRS, Atelier Lexicographie
B.P. 8. 7, rue Guy Môquet, 94801 Villejuif Cedex
http://celia.fr/∼delprat
[♯]Université de Franche-Comté, Laseldi 30, rue Mégevand, 25030 Besançon Cedex
alain.lelu@univ-fcomte.fr,
http://laseldi.univ-fcomte.fr/php/accueil.php?cas=14&id=36
[♯♯]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 LaTeX avec le système mayaTeX, 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 en Amérique centrale pendant plus de 13 siècles avant de disparaître sous l'inquisition espagnole au début du 15[e] siècle, nous est parvenue au travers de riches inscriptions sur des monuments, des céramiques et trois almanachs divinatoires. Toutefois, il nous faut faire face à la contrainte drastique d'un volume

faible de textes disponibles : trois manuscrits et quelques milliers d'inscriptions courtes découvertes. Leur déchiffrement bénéficie cependant 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 glyphes est établie de façon certaine pour plus d'un cinquième, et plausible pour une bonne moitié. 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. Le codex de Dresde a été saisi à l'aide du package mayaTEX (Delprat et Orevkov, 2007).

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 :

- 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,
- floue : chaque unité statistique, texte élémentaire ou attribut, se voit attribuer une valeur de centralité (pour parler comme Rosh : *typicality* pour ce qui concerne les textes, *cue validity* pour ce qui concerne les attributs) 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 [1],
- 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 LaTeX

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.

1. Notre démarche se démarque sur ce point du «fuzzy clustering» (Bezdek et Dunn, 1975) : 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.

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 : 

063.403/325 044.063/153 056.202.030 117.147/063

Certains affixes dans des compositions complexes, peuvent se comporter également comme un élément central fixe autour duquel les autres affixes s'orientent. Des affixes peuvent aussi s'inscrire à l'intérieur d'un élément central, au lieu de lui être simplement accolé, et constituer ainsi des ligatures.

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.

Le codage informatique de textes mayas suppose de se livrer au préalable à 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).

Saisie et composition informatique des glyphes mayas sous mayaTEX Un outil informatique original de saisie et édition de textes hiéroglyphiques mayas mayaTEX (Delprat et Orevkov, 2007) est utilisé pour la composition de la paléographie et des lexiques intégrés à la thèse. Développé sous TEX , il comprend 3 polices hiéroglyphiques mayas. 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. Les parenthèses "(" et ")" permettent de constituer un sous-ensemble de glyphes  (154.123) sur lequel agit l'opérateur "." ou "/", par exemple :  (154.123)/177.504 .

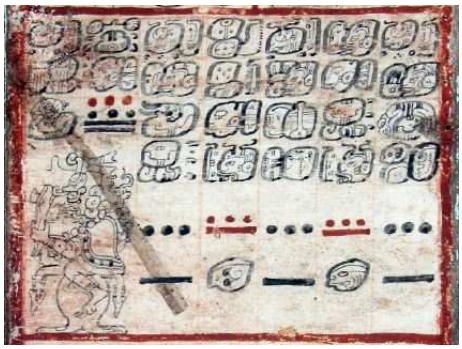
Un ou plusieurs éléments affixe ou central peuvent s'inscrire à l'intérieur d'un élément central géométrique ou, le plus souvent, d'une tête au lieu de lui être simplement accolé ;

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 023/023 sont inscrits au centre de 369. Les notations 359<023/023>et 373 sont équivalentes, elles afficheront la même ligature.

Les éléments inscrits en ligature peuvent être placés au milieu, en partie supérieure, inférieure, gauche ou droite d'un élément central. Les opérateurs habituels de composition " . " et "/" s'appliquent, complétés par "< > ".

3 Notre corpus, le codex de Dresde

Notre corpus est constitué des 74 folios du Codex de Dresde, l'un des trois seuls manuscrits mayas qui nous soient parvenus et datant probablement du 15e siècle, codé selon les numéros du catalogue de (Évréïnov et al., 1969) complété pour mayaTEX. Le corpus est scindé en 212 segments qui correspondent chacun en règle générale à une partie supérieure, centrale ou inférieure de folio, telles que définies par les scribes mayas en général séparées par une ligne rouge. Chacune comprend 1 à 7 blocs de cartouches hiéroglyphiques, formant autant de phrases.



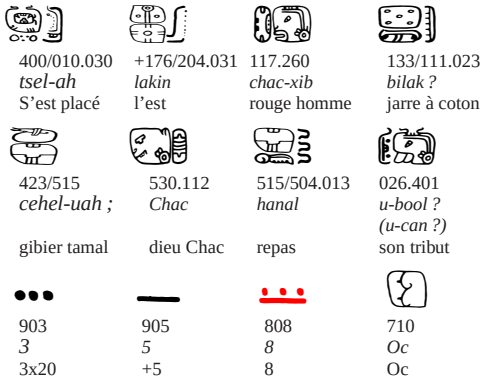


FIG. 1 – *Cliché 30b segment médian du folio 30 du codex de Dresde*

La Figure 1 donne la reproduction du segment numéroté Page 30b avec la paléographie hiéroglyphique correspondante au 2e bloc de texte 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 orients. 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.*

Le codex de Dresde est constitué d'un ensemble de 76 almanachs répartis en 5 types principaux :

– almanachs divinatoires du calendrier *tzolkin* ou cycle de 260 jours consacrés à diverses divinités,
– prophéties de l'année solaire *haab* de 360 jours plus 5 jours intercalaires "sans nom" 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,
– almanachs des cérémonies de la nouvelle année et du déluge associé au cycle *katun* de 52 ans.

Ces textes sont en géneral indépendants les uns des autres et non les chapitres successifs d'un livre occidental à lire d'un début jusqu'à une fin.

Les folios numérotés par (Förstemann, 1880) 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 segments de folio tels que définis par les scribes mayas, en général séparés par une ligne rouge. Chacun comprend 1 à 7 blocs de texte qui correspond à la phrase maya.

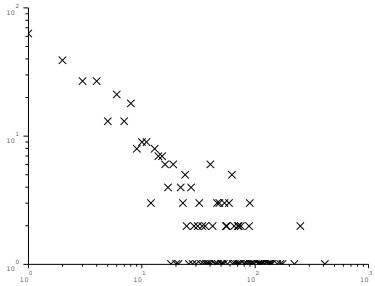


FIG. 2 – *Répétitions des glyphes dans le corpus sans décomposition des ligatures, en coordonnées log-log (abscisses : occurrences de chaque glyphe dans le corpus, ordonnées : nombre de répétitions de ces occurrences).*

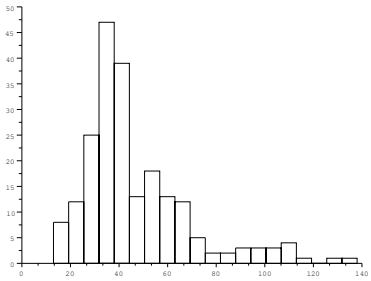


FIG. 3 – *Histogramme du nombre de glyphes sur les segments du corpus (abscisses : longueur des sections, ordonnées : effectifs des classes)*

Du point de vue quantitatif, le corpus comporte 9938 occurrences de glyphes, parmi lesquels 411 distincts [2]. La répartition des glyphes dans le corpus a une allure zipfienne classique (cf Figure 2), d'allure linéaire (en coordonnées log-log) caractéristique d'une loi de puissance. La répartition du nombre de glyphes par segment de texte a elle aussi une allure classique, binomiale (cf Figure 3).

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

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 de segment de texte : Page 5b

 ....  `322/023.030 111.274 904 806` => phrase

 `111.274` = cartouche  `111`= glyphe  `274`= glyphe  `806` = cartouche

Pourquoi des n-grammes ? Choix de n *AL : des n-grammes versus des mots*

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 maya : 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 avions 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.

Exemple :	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 : $\mathbf{z}_{t1} : \left\{\sqrt{\frac{x_{it_1}}{x_{.t_1}}}\right\}$; $\mathbf{z}_{t2} : \left\{\sqrt{\frac{x_{it_2}}{x_{.t_2}}}\right\}$

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 :
$Dd(t_1, t_2) = ||\mathbf{z}_{t1} - \mathbf{z}_{t2}||$

où $||\mathbf{x}||$ 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 :

- 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, ...
- 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 $\mathbf{x}_p$:

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

- Elle est particulièrement adaptée aux «données directionnelles» (Banerjee et al., 2005) que sont les données textuelles, pour lesquelles seuls sont pertinents les angles entre vecteurs, et non leurs longueurs,
- Elle est rapide à calculer dans le cas des données textuelles, où les vecteurs $\mathbf{z}_t$ sont très creux,
- 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'hypersphere. 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 $\left\{x_{i.}^{-\frac{1}{2}}\right\}$ (respectivement $\mathbf{D}_c^{-\frac{1}{2}}$ avec $\left\{x_{.j}^{-\frac{1}{2}}\right\}$), la décomposition aux valeurs singulières (SVD) de $\mathbf{X}^{\frac{1}{2}} = \left\{x_{ij}^{\frac{1}{2}}\right\}$ s'écrit : $\mathbf{X}^{\frac{1}{2}} = \mathbf{UDV}'$ 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 : $\begin{array}{l}\mathbf{F} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{UD} \\ \mathbf{G} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{VD}\end{array}$

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 :

$\mathbf{Q} = \mathbf{U}_{ca}\mathbf{D}_{ca}\mathbf{V}'_{ca}$

où les facteurs AFC s'écrivent : $\begin{array}{l}\mathbf{F}_{ca} = x_{..}^{\frac{1}{2}}\mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}_{ca}\mathbf{D}_{ca} \\ \mathbf{G}_{ca} = x_{..}^{\frac{1}{2}}\mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}_{ca}\mathbf{D}_{ca}\end{array}$

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, definissent les indices de centralité de ces documents [3], 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

3. Il découle directement des propriétés des vecteurs propres que le premier vecteur propre de la table des cosinus

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.

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 2e 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

En fait 2 plans d'analyse :
- le tri thématique des docs dans les classes est vérifié,
- la formation de classes de contenu selon les expressions mayas est précisée par la décomposition des ligatures.

5.1 Sans information exogène

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. Les ligatures de glyphes élémentaires n'ont pas été décomposées en leurs éléments constitutifs et apparaissent comme des glyphes à part entière.

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

$\mathbf{X}^{\frac{1}{2}} \mathbf{D}_c^{-1} \mathbf{X}^{\frac{1}{2}\prime}$ 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.

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  .

La table 2 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 sections correspondantes 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.

BD : Table des 10 classes et leur signification. Complétude de l'ensemble des classes vs. la connaissance experte qu'on a du corpus.

5.2 Avec information exogène = le développé des glyphes-ligatures

BD : Définition, intérêt.
AL : changements des caractéristiques générales, paramétrage.
BD : Les résultats. Comparaison /expérience 1.
MH : Reste à faire : N-grammes à cheval entre 2 cartouches successifs

6 Conclusion et perspectives

AL+BD : résumé des conclusions sur le domaine d'application et la validité de l'assemblage de techniques qui les a permises. choix éventuel d'un sous-corpus, extension des corpus : au codex de Madrid et Paris, inscriptions sur céramiques

Le codage du Codex de Dresde sous mayaTEX 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.

Beaucoup de variantes restent à explorer :

- changer le paramètre N, combiner les 1-grammes et les 2-grammes, ...
- décrire les textes par des "pseudo-N-grammes" ou *triggers* (Lau, 1993), relâchant la contrainte de stricte consécutivité des N-grammes ; ou encore par des motifs de glyphes [4] (Cadot et Lelu, 2007), qui relâcheraient la contrainte de séquentialité au sein des cartouches, tout en préservant une séquentialité inter-cartouches.
- explorer d'autres niveaux de granularité des unités statistiques de textes : phrases, cartouches, ... éventuellement sur un sous-corpus homogène, comme le calendrier des phases de Vénus.
- étendre le corpus aux deux autres codex, voire aux inscriptions murales et poteries.

AL+BD : Perspectives : autres niveaux de granularité (phrases ? cartouche ?)

Nos premières expériences nous encouragent à poursuivre dans cette voie.

Références

Banerjee, A., I. Dhillon, J. Ghosh, et S. Sra (2005). Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research (JMLR) 6*, 1–39.

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

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 15e siècle*. Thèse de doctorat, Institut National des Langues et Civilisations Orientales, Paris. (thèse en cours).

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

Delprat, B. et S. Orevkov (23-27 juillet 2007). mayaTEX} – 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 *JADT 1998 : Actes des 4esJournées Internationales d'Analyse Statistique des données Textuelles*.

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. *Questiio (Quaderns d'Estadistica i Investigacio Operativa) 19*, 23–63.

Renyi, 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 drevhikh 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)	*baktun / pictun*	cycle de 20 x20 x18 x20 = 144.000 jours ou 400 ans	D61AB, D69A

Rang	Expression	Translittération en yucatèque colonial	Signification et traduction	Segments correspondants
2	 069/(-505.322.-505) 069c/(-505.274.-505)	*pawah thul pawah cizin*	divinité lapin divinité mort [incertain]	D61AB, D69A
3	 173/112	*uinic*	homme, cycle de 20 jours	D61AB, D69A
4	 023.153.023)/220	*katun*	cycle de 20X18x20=7.200 jours ou 20 ans	D61AB, D69A
5	 220/009	*tun/haab*	cycle de 18x20=360 jours, année solaire (360+5 jours intercalaires “sans nom”)	D61AB, D69A
6	 105/155.030	*pat otoch-ah / pat-ah / kat-ah*	mis en la maison/former	D52b, D61AB, D69A
7	 054.212	*och ixim/ha’*	entrer [dans le] maïs/l’eau	D31a, D61AB
8	 056/212	*ti ixim/ha’*	dans le maïs/l’eau	D51a, D52a, D61AB, D69B
9	 060/ ?705.030	*o chicchan-ah*	[non compris]	D61AB
10	 204/031	*kin*	jour	D61AB
11	 245.235	*yax Ahau*	le Seigneur vert	D61AB, D69A
12	 804.700	*chan ahau*	le 4 Ahau [date du calendrier Tzolkin de 260 jours]	D31a, D51b, D69A, D70b
13	 411/515	*cumku*	mois maya Cumku	D31a
14	 075.300/031	*yoon / yoon kin*	parents du soleil [incertain]	D61AB, D69A

TABLE 2: *Expressions maya de la Classe 1 par centralité. Codex de Dresde codé sans décomposition des ligatures en glyphes élémentaires.*

Codage et classification non supervisée d'un corpus Maya

Summary

These instructions for preparing RNTI articles with LaTeX should be strictly respected to ensure a homogenous presentation of all articles. Thank you for not modifying the formatting rules. This abstract should not exceed 150 words.