Analyse textuelle de manuscrits mayas et égyptiens : apports d'un codage par n-grammes, et de représentations multidimensionnelles graduées

Bruno Delprat, Martine Cadot, Alain Lelu

Analyse textuelle de manuscrits mayas et égyptiens : apports d'un codage par ngrammes, et de représentations multidimensionnelles graduées

Bruno Delprat[1], Martine Cadot[2], Alain Lelu[3]

[1]Université de Grenoble– mail@mail

[2]LORIA – martine.cadot@loria.fr

[3]retraité, Université de Franche-comté – alelu@orange.fr

Abstract

Texts in logosyllabic scripts are generally written without separators between lexical units. For ancient logosyllabic writings, we propose to explore methods without prior tokenization, adapted to small corpora. We present here a comparative analysis of literary and religious texts, Egyptian tale of the Shipwrecked Sailor, and the only three available Mayan manuscripts, using their representation in n-grams of elementary signs, visualized through LaTeX with mayaTeX, and their processing by Correspondence Analysis and graded unsupervised classification (Axial K-Means and Non-negative Matrix Factorization). The statistical units are sentences and text sections: Egyptian rubrics and Mayan almanacs. We identify intra- and inter-text features, characteristic of the narrative structures in these literary corpora, such as parallelism and *mise en abyme*. The groupings identified on nuanced axes and their correspondences within original texts make it possible to clarify the meaning of certain poorly understood passages, by situating them in contexts easier to interpret.

Keywords:logosyllabic scripts; Maya; Egyptian; n-grams; correspondence analysis; CA;axial k-means; KMA; non-negative matrix factorization; NMF; intrinsic dimension, Monte-Carlo simulations,Tournebool algorithm

Résumé

Les textes en écritures logosyllabiques sont de façon générale écrits sans séparateurs entre les unités lexicales. Pour les écritures logosyllabiques anciennes, on se propose d'explorer des méthodes sans tokenisation préalable, adaptées à de petits corpus. Nous présentons ici une analyse comparative de textes littéraires et religieux, d'une part égyptien du Conte du naufragé, et d'autre part mayas des trois seuls manuscrits mayas disponibles, utilisant leur représentation en n-grammes de signes élémentaires, visualisés sous LaTeX avec mayaTeX, et leur traitement par Analyse Factorielle des Correspondances et classification non supervisée graduée (K-Moyennes Axiales et Non-negative Matrix Factorization). Les unités statistiques sont les phrases et les sections de texte : rubriques égyptiennes et almanachs mayas. Nous en dégageons des manifestations, intra- et inter-textes, caractéristiques des structures narratives dans ces corpus littéraires, comme le parallélisme et la mise en abîme. Les regroupements dégagés sur des axes nuancés et leur report dans le texte original permettent d'éclairer la signification de certains passages peu compris, en les resituant dans des contextes interprétables.

Mots clés : écritures logosyllabiques ; Maya ; Egyptien ; n-grammes ; analyse factorielle des correspondances; AFC ; k-moyennes axiales; KMA ; non-negative matrix factorization ; NMF ; dimension intrinsèque, simulations de Monte-Carlo, algorithme TourneBool

1. Introduction : Traiter les écritures anciennes

Les langues anciennes à écriture logosyllabiques, comme l'égyptien et le maya, incomplètement connues ou déchiffrées, aux orthographes très variables, et sont documentées par des textes monumentaux, sur céramiques et manuscrits disponibles en volumes bien plus

restreints que pour celles vivantes aujourd'hui, principalement le chinois et le japonais. Ces dernières font actuellement l'objet d'abondantes analyses textuelles informatisées avec segmentation préalable (HanLP, 2021), dite tokenisation, fondée sur les nombreux dictionnaires et études linguistiques de ces langues vivantes.

Le nombre actuellement disponible d'inscriptions égyptiennes monumentales, peintes ou gravées sur des parois, meubles et sarcophages ou encore de rouleaux de papyrus est considérable grâce au climat très sec de l'Egypte et aux chambres mortuaires en pierre, propices à la conservation. Les textes les plus longs sont ceux dits des « Livres des morts » accompagnant un riche défunt, comme le papyrus du livre des morts de Nouou (Budge, 1899) qui rassemble environ 93.000 hiéroglyphes en 131 chapitres, des instructions indépendantes à accomplir face au dieux lors de sa navigation éternelle. Les récits littéraires, comme celui de notre corpus d'étude, ou historiques rassemblent aux alentours de 5000 hiéroglyphes.

Comme dans l'antiquité chinoise, les grands romans ou épopées sont absents ; les textes sont courts. Pour son déchiffrement au début du 19e siècle, Champollion s'est appuyé sur les textes de stèles, à quelque 1000 hiéroglyphes. Le cerveau humain (et celui des bébés) déduit rapidement les règles du langage d'un nombre limité de situations rencontrées ; sans recours à l'apprentissage sur de grands volumes de parole ou de textes.

L'écriture logosyllabique des Mayas utilisée sur plus de 13 siècles, est parvenue par de riches inscriptions qui constituent néanmoins un volume plus faible de textes disponibles: trois manuscrits almanachs divinatoires (ensemble env. 24.600 glyphes), quelques milliers d'inscriptions courtes inscrites en caractères cursifs sur poteries et peintures murales, et en glyphes monumentaux sur des stèles et monuments (de l'ordre de 500 glyphes chaque), avec de très nombreuses variantes graphiques équivalentes.

Nous vivons l'essor des « Big Data », susceptibles de traiter des millions, voire des milliards de textes, par des méthodes qui respectent la dimension temporelle du discours, et pour certaines [Vanni et al. 2023] mettent en lumière le « pourquoi » des résultats, contrairement au fonctionnement opaque des autres. Mais les besoins d'avancées de la recherche sur les « small data » sont toujours prégnants, car dans certains secteurs, comme précisément l'étude des textes antiques, on ne dispose que de «Small Data ». Dans ce contexte la puissance des matériels informatiques permet de reculer les limites des méthodes énumératives, de faire progresser l'optimisation de méthodes convergeant vers des optima locaux, comme les classifications à centres mobiles : là où on se contentait d'une vingtaine de passages avec des graines d'initialisation différentes, on peut en envisager des milliers, voire dizaines de milliers qui nous rapprochent encore de l'optimum optimorum. Créer des centaines ou milliers de clones aléatoires d'un tableau de données permet d'en extraire la dimension intrinsèque, comme on le verra plus loin, et même si cette dimension dépasse les capacités humaines d'interprétation, elle ouvre la voie à d'autres traitements gourmands en calcul.

2. Etat de l'art et choix pour les écritures logo-syllabiques

2.1. Caractéristiques des écritures égyptienne et maya

Les hiéroglyphes égyptiens tout comme les caractères chinois, dérivent de signes à l'origine figuratifs longs à tracer qui ont subsisté avec de nombreuses variantes pour les inscriptions monumentales tout au long des 3600 ans du règne des pharaons. Les signes, plats, verticaux ou carrés sont assemblés pour occuper au mieux tout l'espace d'un « quadrat », le mot polysyllabique pouvant s'étaler sur plusieurs quadrats. Parallèlement dès la haute antiquité, les

scribes ont établi une écriture cursive aux tracés filiformes simples et peu figuratifs, efficace pour écrire à l'encre sur papyrus. Dite « hiératique », c'est celle du Conte du Naufragé.

L'égyptien ancien est une langue à flexions. Il est doté d'une riche morphologie, avec de fréquents suffixes grammaticaux et signes déterminants sémantiques attachés aux mots, ainsi qu'un jeu de particules grammaticales indépendantes. Chaque élément grammatical ou déterminant est représenté généralement en 1 signe d'écriture ; les mots racines en 2 à 7 signes. Le signe est polyvalent, selon le contexte il peut se lire comme : logogramme, à partir de sa valeur figurative la plus ancienne ; signe phonétique représentant un son ou une syllabe dans un mot ; flexion grammaticale sur sa base phonétique ; ou encore déterminant sémantique ou catégoriel adjoint en fin de mot qui ne se prononce pas. L'écriture égyptienne est alors composite, comme l'est celle du japonais, associant dans la même phrase des mots écrits selon des principes différents: logogramme + déterminant, phonétique(s) + déterminant, logogramme + élément phonétique + déterminant, signes à valeur phonétique seuls.

Dès l'Egypte antique, des dictionnaires ont existé, groupant les hiéroglyphes par familles de représentations graphiques (Ciel, terre et eau ; l'Homme et ses occupations ; les Oiseaux ;...). L'égyptologue Sir Gardiner a repris ce système en attribuant des lettres (N ; A ; G ; ...) à ces catégories suivies d'un numéro séquentiel pour les signes. Le "Manuel de codage" (Buurman et al., 1988) a normalisé pour l'informatique un codage Gardiner étendu, ici utilisé.

Le signe d'écriture élémentaire maya est le glyphe, élément central presque carré, ou affixes plats qui s'organisent en tournant et par symétries, selon une règle déterminée, autour des éléments centraux. Les éléments centraux sont le plus souvent des logogrammes correspondant à un morphème ou mot, qui par effet "rébus" servent souvent d'élément syllabique. Les affixes ont avant tout une valeur phonétique syllabique. 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 disponible dans la page d'almanach, le scribe pouvait entasser ou étaler 1 à 8 éléments de base dans les cartouches, les groupant par expressions signifiantes, veillant à une belle mise en page. Le découpage en cartouche des scribes est une bonne approximation à celui des mots ou expressions mayas, incluant les préfixes ou post-fixes grammaticaux. A la différence des textes égyptiens, les textes mayas présentent un état condensé de la langue, dépouillé des flexions et particules grammaticales non indispensables, ce que fait aussi le chinois classique.

Dans les années 1960 des cryptologues et historiens soviétiques (Évréïnov et al., 1961-69) ont entrepris un codage informatique sur 3 chiffres des signes manuscrits de l'écriture qui comprend environ 520 glyphes. Nous l'avons utilisé dans notre corpus des codex mayas.

2.2. Prétraitements des textes

L'écriture hiératique égyptienne fusionne sous la même graphie cursive simplifiée quelques hiéroglyphes à la lecture distincte. Inversement, des signes ont plusieurs graphies hiératiques, selon qu'ils sont comprimés associés à d'autres ou occupent seul le quadrat. Nous avons réduit 13 de ces signes à une seule graphie dans notre corpus. Dans le cas du maya également, des glyphes fréquents présentent des variantes graphiques interchangeables au choix du scribe, nous a amenant à assimiler 98 de ces variantes dans notre catalogue des signes. De plus, certains glyphes élémentaires s'infixent dans un autre glyphe, au lieu de se positionner sur un de ses bords, mais pour des raisons pratiques, la police de caractères les traite souvent graphiquement comme un signe. Nous les avons décomposé en leurs éléments logiques dans notre corpus, ce

qui réduit encore de 90 les glyphes élémentaires différents considéré dans nos traitements sur le corpus. Le catalogue des glyphes mayas est alors réduit à 303 glyphes.

Pour éviter que les termes mayas mono-glyphes ou égyptiens bi-glyphes ne soient éliminés lors des traitements sur le corpus, nous leur avons infixé le code vide “666”. Formant ainsi artificiellement des bigrammes ou trigrammes, ils sont alors pris en compte. De façon similaire, un code vide “777” peut être utilisé comme séparateur de mot logique égyptien ou de cartouche maya, ce qui permet d’exclure des analyses les n-grammes « à cheval » sur plusieurs mots ou cartouches dans une partie de nos traitements.

2.3. Visualisation des résultats hiéroglyphiques sous LaTeX

Une étape préliminaire a été la constitution et vérification par rapport aux reproductions des manuscrits originaux des paléographies hiéroglyphiques informatiques de nos deux corpus égyptien et maya sous LaTeX, à l'aide du système mayaTeX (Delprat et Orevkov, 2007), un outil informatique original développé sous TeX de saisie et édition à l'origine de textes hiéroglyphiques mayas. Il permet de composer les signes hiéroglyphiques de base dans le cartouche maya tout comme le quadrat égyptien. Les deux principaux opérateurs de composition des signes sont le point "." qui associe deux signes en les juxtaposant latéralement dans le sens gauche-droite, et l'opérateur ":" ou "/" pour la juxtaposition verticale de haut en bas. On peut grouper des sous-ensembles de signes à l'aide de parenthèses. Des polices de caractères mayas ont été élaborées, ici sert la police Codex , auquelles se sont ajoutées plusieurs polices égyptiennes dont ici: Imprimerie nationale de style silhouette dérivée des inscriptions monumentales, et Ebers dérivée des signes hiératiques.

2.4. Représentation

L'état de l'art des méthodes d'analyse textuelle "non-Big Data" ne permet pas de traiter pleinement la continuité des textes. Pour l'approcher le moins mal possible, nous avons choisi de les coder au moyen de n-grammes de signes élémentaires : trigrammes pour l'égyptien, bigrammes pour le maya. Deux niveaux de granularité d'analyse (Lelu et Roussanaly, 2014) ont été sélectionnés : celui de la phrase (égyptien), du bloc (maya), et celui moins fin mais respectueux des catégories mentales des scribes, de la rubrique (égyptien) ou de l'almanach (maya). Nous recherchions la possibilité de nuancer l'interprétation : certains éléments sont plus centraux que d'autres dans une classe, certains, importants dans une classe, peuvent être également importants dans d'autres contextes. La représentation par axes de classes, où se projettent non seulement les éléments d'une classe, mais aussi ceux des autres classes, propre aux KMA et à la NMF, s'est imposée à cet égard. Un avantage supplémentaire de la représentation axiale est de pouvoir comparer sur un même ensemble deux classifications issues de descripteurs différents, par exemple des publications décrites par des mots vs. des liens de citation (Lelu et al., 2013). Un autre choix délicat est celui du nombre de classes. Si notre procédure TourneBool, qu'on décrira plus loin, permet de fixer de façon rigoureuse un nombre d’axes-classes linéairement indépendants (Lelu A. et Cadot M., 2011) à ne pas dépasser - ici 21 pour le classement des trigrammes égyptiens - l'ordre de grandeur en est excessif pour les facultés humaines. Six classes nous ont paru un compromis raisonnable entre finesse d'analyse et charge mentale d'interprétation pour l’égyptien ; et 15 classes dans le cas des 3 codex mayas.

3. Principes des traitements

On ne présentera pas l'Analyse Factorielle des Correspondances, utilisée sans discontinuer depuis des décennies en analyse des données textuelles (Lebart et Salem, 1994). Sa propriété

essentielle d'équivalence distributionnelle - fusionner 2 lignes ou colonnes de même profil relatif ne change pas l'analyse - est partagée par l'Analyse Factorielle Sphérique (AFS) (Domengès et Volle 1979 ; Hallab et al., 2010). Alors que l'AFC transforme le nuage des données d'origine à n dimensions en un plan "simplexe étiré" (Greenacre et Hastie 1987) à n-1 dimensions, orthogonal à un vecteur unité, lequel constitue son premier facteur, trivial par définition, l'AFS dans sa version "distance au tableau nul" dispose ce nuage à la surface de la sphère unité, et son premier facteur s'interprète comme un indicateur de centralité des points par rapport à cet axe, sur lequel leurs projections se disposent à proportion de leur caractère typique dans ce nouveau nuage de points - celles-ci de 0 à 1 si les valeurs du tableau d'origine sont positives. Les K-Moyennes-Axiales (KMA) (Lelu, 2008) utilisent cette représentation et répartissent les points à la surface de la sphère unité selon un "pavage de Voronoï" autour des axes de classe : après initialisation au hasard des K axes de classes, son algorithme à centres mobiles converge vers un optimum local de la fonction objectif. Celle-ci mesure la similarité du tableau reconstitué à celui d'origine, minimisant leur écart quadratique - dans le cas de l'AFS et des KMA : la part de la somme des valeurs du tableau d'origine qu'elle reconstitue.

La méthode de décomposition matricielle NMF (Non-negative Matrix Factorization) (Lee et Seung, 1999) décompose la matrice A (m lignes, n colonnes) en produit des deux matrices W et H de dimensions respectives (m,k) et (k,n) sous la contrainte de non-négativité des éléments de W et H, k étant donné. Elle aussi aboutit à k axes obliques pointant dans les directions de zones de densité importante du nuage de points. Pour faire court, la décomposition aux valeurs singulières, sous-jacente à l'Analyse en Composantes Principales (ACP) et à l'AFC, maximise la précision de reconstitution, alors que la NMF, dénuée de la contrainte d'othogonalité des axes, maximise l'interprétabilité. On peut tirer une classification des lignes de A à partir des valeurs maximales de leurs projections sur les k axes définis par H. Sa formulation de base, appliquée aux racines carrées des valeurs du tableau d'origine, minimise comme les KMA l'écart quadratique avec ce tableau. Nous comparerons plus loin sur un exemple KMA et NMF, deux représentations très proches en ce qu'elles présentent des axes de classes gradués, et non des appartenances en tout ou rien comme les méthodes de clustering classiques.

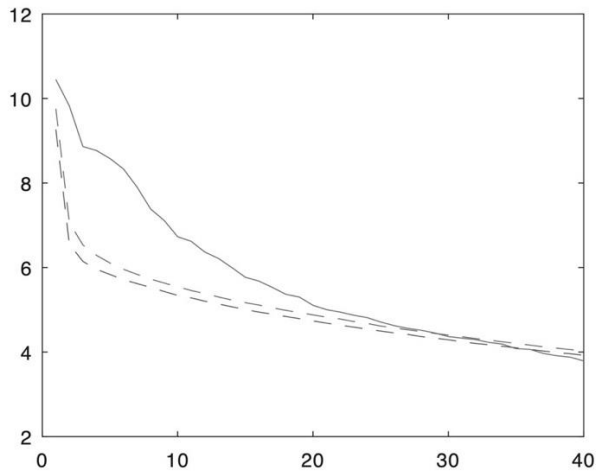

Figure 1 : Eboulis des valeurs propres du tableau 577 trigrammes × 143 phrases. En pointillés : intervalle de confiance à 95% établi par simulations Monte Carlo. Conte égyptien Naufragé

Combien de classes pertinentes peut-on tirer d'un tableau de données ? Quand ces classes regroupent des profils relatifs linéairement indépendants – constat empirique - on aura autant de dimensions distinctes de l'espace que de classes. On appelera dimension intrinsèque le nombre maximal de ces dimensions, nécessairement inférieure au minimum des nombres de lignes et de colonnes. Dans Cadot 2005 et (Cadot et Lelu, 2009) on a développé une méthode de Monte Carlo produisant à partir d'un tableau binaire (le caractère très clairsemé des tableaux de données textuelles permet en pratique de les assimiler à leurs contreparties binaires), via un grand nombre de permutations aléatoires, des tableaux binaires simulés de mêmes sommes marginales que celui d'origine. Les valeurs propres de ces matrices, rangées par ordre décroissant, engendrent un intervalle de confiance au dessus duquel les valeurs propres du tableau d'origine ne sont pas porteuses de bruit, et ont donc du sens (Figure 1). Leur nombre, nous l'appellerons dimension

intrinsèque du tableau. Il fixe le nombre maximal de classes, au sens défini plus haut, autrement dit d'axes ou facteurs, orthogonaux ou obliques, qu'on peut en extraire.

4. Application des AFC / KMA à des corpus anciens

4.1. Le Conte égyptien du naufragé

Le papyrus Ermitage N° 1115 du Moyen empire bien conservé et complet, comprend la seule version retrouvée de ce conte, vraisemblablement écrit au début de la XIIième dynastie au 20ième siècle av.J.-C. où les expéditions minières renforcent l'Égypte sur ses voisins.

La disposition des signes dans les quadrats suit la composition en écriture hiératique du papyrus ; photographies et paléographie ont été publiées par Golénischeff (1913). Les phrases sont écrites toutes à la suite sans respirations, exceptés 20 segments notés par le scribe, avec les rubriques à l'encre rouge de leurs premiers mots. Ce texte comporte un total de 3417 hiéroglyphes, découpé en 143 phrases et mots logiques par l'égyptologue Poe (2008). Dans de rares cas, comme le *conte du Prince prédestiné* (papyrus Harris n°500) (Maspéro, 1879-86), les anciens égyptiens ont marqué de points rouges la fin des mots au dessus du texte, comme aide à la lecture.

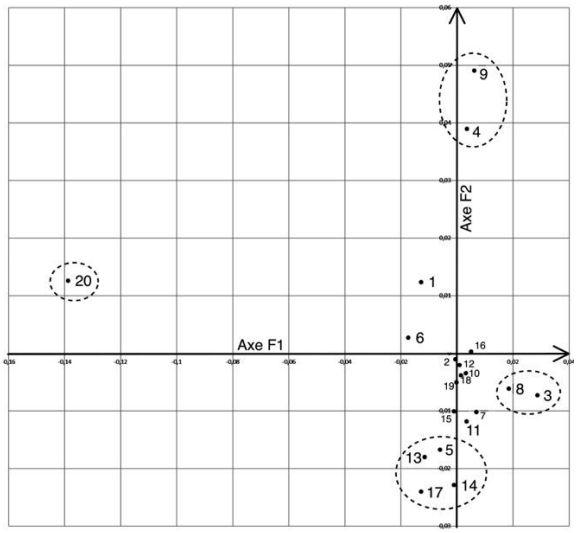

Figure 2 : Diagramme des axes F1 et F2 par analyse AFC des rubriques du corpus égyptien

Les groupements par analyse AFC selon l'axe F2 des rubriques (Figure 2) permet de dégager les associations suivantes:

- Rub. 17, 14, 13 et 5 : sections du texte contenant des énumérations de produits et animaux (les autres rubriques n'en contiennent pas)
- Rub. 4 et 9 : sections sur la tempête et en conséquence le naufrage (on n'en parle pas ailleurs)
- Rub. 3 et 8 : sections décrivant le bateau et son équipage (on n'en parle pas ailleurs)
- Rub 20 : genre de notice bibliographique, écrit dans un style différent de celui du récit.

Dans un texte littéraire comme celui-ci, il sera rare de trouver une réelle opposition entre des sections du récit. Les constructions y sont nuancées, plus subtiles que dans un texte argumentatif, où se dégagerait de la structure des oppositions nettes entre segments.

Ces trois groupements de rubriques correspondent aussi à l'effet de parallélisme dans le récit, plus précisément c'est la mise en abyme du récit du serpent roi/dieu imbriqué à l'intérieur de celui du naufragé. L'étude rhétorique et quantitative des tutoriels en réalité virtuelle par AFC (Dozo et Barnabé, 2022) montre que le « hub » de ces jeux instaure une mise en abyme, comme médiation régulièrement employée pour apprendre au joueur à « prendre » le jeu. Nos premiers résultats de l'AFC mettent en évidence aussi cette mise en abyme, utilisée par le scribe auteur comme médiation ludique et didactique : géographie économique de la Mer rouge, techniques de navigation et phénomènes climatiques. Ainsi, l'AFC sur les rubriques renseigne sur les intentions derrière le récit, dégage des synergies et non des oppositions.

On a avec ce double récit, des rubriques situées dans des temporalités différentes (présent/passé ; domaine humain/divin). La différence entre plans temporels est donnée par l'emploi des temps, de la voix, de la personne, des déterminants sémantiques.

En langue égyptienne, ces éléments grammaticaux sont rendus par des particules généralement écrites en un seul signe hiéroglyphique. Pour le récit Nedjma de Kateb Yacine, Chiali (2013) a effectué le traitement par l'AFC de la structure temporelle par les verbes du français, s'agissant de voir comment l'AFC valide d'une démarche lexico-statistique, qui permet de traquer la variance temporelle et l'examen de certaines récurrences. Une étude AFC des contextes d'emploi des éléments grammaticaux, croisant et quantifiant parallélisme et temporalités, permet l'étude des procédés littéraires dans un texte égyptien à finalité aussi largement informative.

Une analyse par KMA (optimisation sur 1000 passages) sur les 577 trigrammes pleins de signes hiéroglyphiques en fonction des 143 phrases, sans découpage en mots logiques a été réalisée, avec appartenance d'un trigramme à plusieurs classes si sa contribution y est > 0,8x la contribution maximale dans les classes. Il en résulte que 42 bigrammes, soit 8% sont attribués au moins à 2 classes. Ces trigrammes multiclasses sont moins nombreux qu'anticipé. En corpus égyptien, l'approche multiclasse apporte quelques nuances par rapport à l'attribution des trigrammes à une classe unique. Trait rassurant, la forte contribution du trigramme phonétique M17.M17.X1 à plusieurs classes, enchâssé à l'intérieur de nombreux mots longs, montre que ces axes se sont construits plutôt par rapport aux racines caractéristiques des mots, et non d'un trait répétitif de morphologie interne.

Dans l'approche découpage en phrases sans séparateurs de mots logiques d'un texte égyptien, les nombreuses flexions et particules grammaticales se retrouvent dans environ 2/3 des trigrammes extraits par KMA, souvent adjointes à la racine des mots, exemples: U2:D4 G1 V31 "tu reverras", ou G17 X1:Z6 V31 "tu mourras". Ces trigrammes ainsi que ceux à cheval sur plusieurs mots caractérisent très largement les traits grammaticaux et structurels du texte, au détriment des associations sémantiques de termes signifiants.

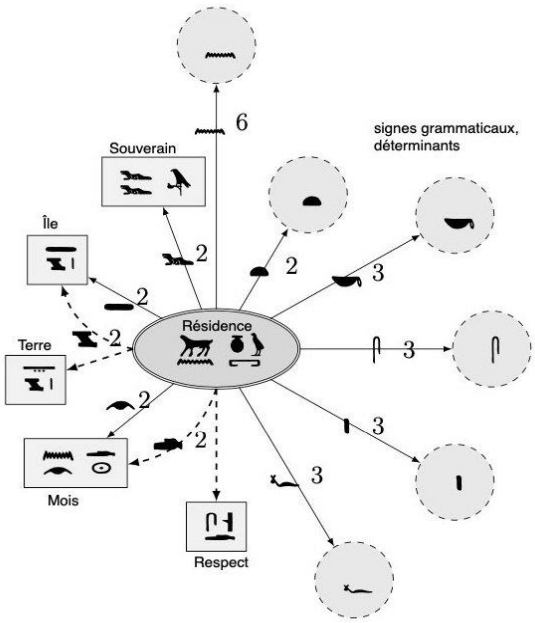

Figure 3 : Voisinage par AFC du terme « résidence ». Conte égyptien du Naufragé

Avec une analyse AFC sur des quadri-grammes du type : "X_Y (2 à 5 trous) P_Q", soit un bigramme complet suivi de 2 à 5 "trous" et d'un autre bigramme complet, nous avons analysé le voisinage du terme F26:N35 (W24.G43):O1 « résidence, chez-soi » (Figure 3). Dans ce diagramme de relations, les termes égyptiens associés sont dans des boites rectangulaires ; les signes grammaticaux ou déterminants dans des cercles pointillés; au centre le terme « résidence » . Le 2ième bigramme complet des 4-grammes est indiqué sur les flèches avec le chiffre de fréquence de l'association.

On observe que dans la totalité des 29 occurrences, les 4-grammes à trous pointent en dehors du mot de départ « résidence, chez-soi », qui s'écrit en 5 signes hiéroglyphiques. 10 des occurrences (35%) mettent en relation

« souverain », « île » et « terre » des thèmes centraux du conte égyptien; mais aussi « mois » et « montrer du respect ». Les 19 autres associations (65%) sont avec des signes à fonction grammaticale ou de déterminant, qui sont adjoints aux mots suivants dans le texte.

On a avec les 4-grammes du type "X_Y (2 à 5 trous) P_Q" une prépondérance d'association d'un mot avec le contexte grammatical dans la suite proche du texte (2/3 des cas), mais aussi pour 1/3 des cas des associations avec d'autres mots. L'étude similaire sur les trigrammes du type "X_Y (2 à 5 trous) P" donne des résultats approchants.

Pour mettre en évidence les associations de mots dans les phrases et rubriques du récit, nous avons repris cette expérimentation KMA phrases logiques sur 6 classes avec séparateurs de mots, pour éviter la prépondérance des phénomènes d'ordre grammaticaux et stylistiques.

L'analyse sur les phrases dégage 5 classes thématiques aux typicités max. élevées > 0,86: Cl. 1: « les produits du pays de Pount », Cl. 2: « observation et orage », Cl. 3: « la houle et les flots », Cl. 5: « prosternation face au dieu serpent », Cl. 6: « lieux de résidence et famille ».

Une seule classe de phrases est caractéristique de sa structure avec une typicité maximale plus faible de 0,62 mais une longue liste de 44 éléments: la Cl. 4: toutes des phrases commençant par P6.D36:N35 « alors, ensuite ». Les thèmes sont significatifs pour une typicité des trigrammes > 0,15 ou 0,25 selon les 6 classes, avec du bruit en fonds de classes. Il n'apparait pas d'autres traits stylistiques ou structurels, comme l'effet de mise en abîme dégagé dans les rubriques sans séparation des mots.

Typicité	Trigramme	Terme	Traduction	Effectif
0,850	P6 D36 N35		alors, ensuite, après	27
0,393	D36 N35 A1			7
0,662	V31 G43 A1		1S parfait	17
0,564	D46 Q3 X1		bateau	11
0,587	Q3 X1 P1			9
0,421	M14 G36 D21		mer (litt. la grande verte)	7
0,421	G36 D21 N36			7
0,413	V4 G1 G43		vague (n.f.)	3
0,413	G1 G43 N35			3
0,413	G43 N35 N35			3
0,374	G17 X1 Z6		mourir	4

Figure 4 : Axe 5 « ensuite*, le naufrage ».*

KMA trigrammes/phrases. Corpus égyptien

Les axes thématiques de trigrammes sont moins définis que ceux sur les phrases: Axe 1: « les flots et leurs mouvements », Axe 4: « bravoure », Axe 5: « ensuite, le naufrage » (Figure 4), Axe 6: « manifestations du respect ». On ne retrouve que 10% de trigrammes à fonction grammaticale avec typicité toujours < 0,6 dans les axes thématiques, alors qu'ils peuplaient avec les trigrammes « à cheval » 60% des classes dans les passages sans séparation des mots.

4.2. Les 3 codex Mayas

Notre corpus est constitué des trois seuls manuscrits mayas (Taladoire, 2012) sauvés des *autodafé* des évangélisateurs ou du climat humide et chaud d'Amérique centrale, préservés dans les bibliothèques de trois villes d'Europe. Ce sont les codex de : Paris (13e siècle), Dresde (15e siècle) et Madrid (16e ou 17e siècle). Présentés comme des paravents écrits et peints sur les 2 faces, ils rassemblent un ensemble d'almanachs de 5 types principaux (Hallab et al., 2010) : 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 *katun*. Ces textes sont en général indépendants et non les chapitres successifs d'un livre à lire du début à la fin. Des blocs de cartouches souvent accompagnés de dates, nombres distances et d'une illustration formaient les

phrases (augures) de ces almanachs. Le corpus analysé compte 957 blocs-phrases, réparties en 137 almanachs.

Un premier passage KMA corpus des codex mayas en 15 classes sur les blocs-phrases sans frontière des mots et avec l'ensemble des nombres et dates associés aux textes augures des almanachs a produit des axes sur les bigrammes dont certains ne contenaient que ces nombres et dates, fortement présents dans les autres axes aussi, devenant difficilement interprétables. Les passages ultérieurs ont été réalisés sur le corpus expurgé des dates et distances inter-dates.

Nos premières expérimentations KMA sur le codex de Dresde (Hallab et al., 2010) étaient fondés sur un découpage en registres supérieur, médian et inférieur des folios. Pour respecter la lecture horizontale transverse des segments de folios, nous les avons ici regroupés en almanachs complets et effectué une analyse KMA (optimisation sur 1000 passages) en 15 Classes de bigrammes (pleins) de glyphes mayas élémentaires en fonction des 137 almanachs et des cartouches de texte, sans tenir compte du niveau inférieur, le découpage en blocs-phrases, ni des bigrammes à cheval sur plusieurs cartouches. Une analyse sur les bigrammes dégage des axes thématiques: Axe 1: « *résonna* et augures » (s'étend sur 60 almanachs), Axe 9 (Figure 5): « les dignitaires et *Kauil*, dieu de la fertilité » (sur 74 almanachs).

Typicité	Bigramme	Terme	Traduction	Effectif
0,9936	154 123		dignitaire, seigneur	155
0,6216	123 306		maître des lieux	31
0,2292	306 504			6
0,5649	123 204		*Kinich Ahau*, Div.G le dieu solaire	17
0,5390	204 504			14
0,2168	026 154		son dignitaire homme jeune	4
0,2113	123 243		perforation du dignitaire	4
0,1481	033 154		perfore le dignitaire	2
0,1753	504 154		*Kinich Ahau*, Div. G le dieu solaire	2
0,1549	123 264		*Kuh*, Div. C le seigneur	2
0,6318	534 666		*Kauil* Div. K, le dieu de la fertilité	25

Figure 5 : Axe 9 « les dignitaires et Kauil ». KMA bigrammes/almanachs corpus maya

5. KMA et état de l'art : comparaison empirique

Dans (Lelu et Cadot, 2019) nous avons confronté les principales méthodes de clustering sur des corpus standards. Comme la NMF est à notre connaissance la seule méthode de classification non-supervisée partageant les mêmes principes de représentation que les KMA, nous confronter à l'état de l'art signifie nous comparer sur un des exemples de tableau présentés plus haut : celui des 577 trigrammes égyptiens dans l'espace des 143 phrases du Conte du Naufragé, tableau de données à la fois petit et équilibré, et qui respecte au mieux la continuité de chaque phrase.

On commence par calculer la dimension intrinsèque de ce tableau, qui fixera une borne utile pour la suite : nous avons engendré avec notre procédure TourneBool 500 versions aléatoires binaires de notre tableau binarisé (celui-ci ne comporte que 49 valeurs autres que zéro et un sur 1756 au total). Celles-ci partagent la même répartition de densité que le tableau de référence, à savoir les mêmes sommes marginales. Les valeurs propres de ces matrices permettent de tracer un "couloir" de confiance à 95%, les 10 plus petites valeurs propres et les 10 plus grandes étant exclues. La Figure 1 montre que c'est pour la 21eme valeur propre (par ordre décroissant) que l'"éboulis" des valeurs propres de la matrice binarisée d'origine atteint ce couloir. Pour comparer les deux méthodes, on procédera en deux temps : d'abord en leur imposant 6 axes ou classes, comme précédemment, puis 21, limite du maximum de reconstitution pertinente des données.

5.1. Avec 6 axes/classes

KMA : on les applique 1000 fois avec des graines d'initialisation de 1 à 1000. Le meilleur passage obtient un taux de reconstitution des données de 416,5/1812=22,98%

NMF : les auteurs du programme utilisé (Kim et Park, 2008) ne proposant pas de graine d'initialisation en paramètre nous avons ajouté cette option. Pour l'équité du test, on transforme ensuite les données pour obtenir une même fonction objectif : $X \rightarrow X^{(1/2)}$, pour 1000 passages également. Le taux de reconstitution est de 23,10%, donc légèrement en faveur de la NMF.

Pour comparer les deux classifications, on utilise de façon classique l'indice NMI (Normalized Mutual Information) (Hubert et Arabie, 1985) et l'Adjusted Rand Index (ARI) (Danon et al., 2005). Ici NMI=0,8249 et ARI=0,7948. Le caractère angulaire des deux méthodes permet de proposer un Indice de Similarité Angulaire Relative (ISAR) : une fois les deux classifications alignées, comme pour l'indice ARI (NMI n'a pas cette contrainte), et le tableau de leurs angles constitué, on calcule les moyennes respectives des angles sur la diagonale et hors diagonale ; alors ISAR =(moyenne hors diag. - moyenne sur diag.) / moyenne hors diag. Ici ISAR= 0,6993. Cet indice angulaire semble cohérent avec les deux autres indices. Leur diverses valeurs montrent une grande proximité qualitative entre les axes trouvés, en cohérence avec le constat que l'on y retrouve le plus souvent les mêmes trigrammes dans un ordre légèrement différent.

5.2. Dans l'espace intrinsèque des données (soit avec 21 axes-classes)

Les KMA aboutissent à 47,61% de reconstitution des données, contre 50,70% avec la factorisation NMF dans l'espace des données transformées comme décrit ci-dessus, avec NMI=0,7992 et ARI=0,6279, valeurs obtenues sur 1000 passages D'où l'hypothèse que, au moins pour ce type de tableau très peu dense, typique des données textuelles, les résultats de la NMF surpassent légèrement ceux des KMA, avec des axes angulairement proches. Mais le temps de calcul des KMA est beaucoup plus rapide : pour 21 classes demandées et une optimisation avec 200 passages, les taux de reconstitution des données sont comparables – resp. 47,36 et 50,47%, mais les KMA prennent 36 fois moins de temps de calcul.

6. Conclusions

Sur des données de taille limitée par nature, issues d'une lignée au total considérable de travaux de déchiffrement, codage et pré-traitement de deux corpus en écritures logosyllabiques anciennes, l'égyptien et le maya, nous avons appliqué l'AFC, puis deux algorithmes de classification non-supervisée aboutissant au même type de représentation nuancée : ces dernières méthodes ajoutent à la création de classes leur représentation de type factoriel.

Lorsque l'on ignore la frontière des mots logiques égyptiens ou des mots mayas, aussi bien l'approche AFC que KMA tendent à révèler principalement la structure du récit avec un découpage en rubriques ou almanachs, ou les rapprochements de construction grammaticale sur un découpage du corpus en phrases. Avec l'emploi des séparateurs de mots ou cartouches, c'est le champ sémantique qui est privilégié avec des classes thématiques qui comportent moins de 10% d'éléments grammaticaux. Dans toutes les analyses KMA, les n-grammes avec typicité < 0,20 environ constituent du bruit en fonds de classes. La séparation ou non des mots apparait donc comme un choix fondamental préalable selon le type de recherche liguistique.

Sur l'exemple traité, force est de constater que les résultats des NMF dépassent de peu ceux des KMA, alors que cette dernière méthode est bien moins gourmande en temps de calcul, toutes choses égales par ailleurs, y compris quand il s'agit d'extraire le maximum d'information pertinente possible, grâce à la notion de dimension intrinsèque d'un tableau de données. Le processus d'extraction de cette dimension et l'optimisation poussée pour les méthodes itératives

utilisées profitent du caractère "Small Data" du problème, inenvisageable à l'échelle des "Big Data". Nous mettrons en ligne sur HAL, sous forme de document associé, les codes Python et Octave utilisés, en licence GPL, ainsi qu'un exemple de chaîne de traitements. Un réglage différencié des paramètres de pré-traitement adapte de façon empirique par tatonnements nos méthodes aux spécificités de différentes écritures logosyllabiques. C'est en autres le cas des n-grammes à "trous" pour l'exploration par AFC de quelques voisinages linguistiques, pour systématiser les contextes structuraux et stylistiques. Plus généralement, l'étude en AFC des contextes d'emploi des particules grammaticales pourrait permettre de classer selon des axes de "temporalité" les différentes rubriques ou sections du récit dans des langues anciennes encore mal connues avec de petits corpus. Des applications similaires par les KMA ou NMF seraient aussi à réaliser. Le travail sur l'ensemble des trois codex mayas réunis pour la première fois sous une forme homogène et exploitable par l'ordinateur ne fait que commencer. Nul doute que les régularités détectées par l'analyse de données textuelle sur l'ensemble pourront nourrir des hypothèses concernant leurs passages dégradés ou incomplets.

Bibliographie

Budge E. A. W. (1899) Facsimiles of the Papyri of Hunefer, Anhai, Kerasher and Netchemet, with suplementary Text from the Papyrus of Nu, Londres: British Museum, Harrison & Sons.

Buurman J., Grimal N., Hainsworth M., Hallof J. et van der Plas D. (1988) Inventaire des signes hiéroglyphiques en vue de leur saisie informatique, Paris: Institut de France, coll. Mémoires de l'Académie des Inscriptions et Belles Lettres, Nouvelle série, t. 8, 215 p.

Cadot M. (2005) A Randomization Test for extracting Robust Association Rules. *3rd world conf. on Computational Statistics & Data Analysis - CSDA 2005*, Limassol, Cyprus. ⟨inria-00337069⟩

Chiali F. Z. (2013) Traitement par l'AFC de la structure temporelle par les verbes, in *Passerelle*, 4(1), Oran : Université Mohamed Ben Ahmed, 88-116

Dozo B.-O. et Barnabé F. (2022) Transposer les grammaires vidéoludiques : une étude rhétorique et quantitative des tutoriels en réalité virtuelle, *Sciences du jeu*, 17, http://journals.openedition.org/sdj/4098

Danon L., Díaz-Guilera A., Duch J. et Arenas A., (2005) Comparing community structure identification, In *Journal of Statistical Mechanics*, 2005 (09): 09008.

Delprat B. et Orevkov S. (2012) MayaPS: Typing Maya Hieroglyphics with TeX/LaTeX. In *TUGboat*, 33(2012), no 3, Boston, 289-294.

Évréïnov E. V., Kosarev Y. G. et Oustinov V. A. Евреинов, Е. В., Косарев, Ю. Г. и Устинов В. А. (1961- 69). Применение электронных вычислительных машин в исследовании письменности древних майя, Utilisation des ordinateurs pour les recherches sur l'écriture des anciens Mayas, 4 vol., Новосибирск: АН СССР, Novossibirsk : Académie des sciences de l'URSS, 1402 p.

Golénischeff V. S. (1913) Les Papyrus Hiératiques No. No. 1115, 1116A et 1116B de L'Ermitage Impérial à St. Pétersbourg, Saint Pétersbourg : Manufacture des papiers de l'État, 79 p.

Greenacre M et Hastie T. (1987) The geometric interpretation of correspondence analysis, *Journal of the American Statistical Association*, vol. 82, no398, 437-447,

Hallab M., Delprat B. et Lelu A. (2010) Codage et classification non supervisée d'un corpus maya. In *Extraction et Gestion de Connaissances - EGC 2010*, Sousse, Tunisie, 573-584. ⟨hal-00435233v2⟩

HanLP (2021) 多语种自然语言处理技术*Multilingual Language Processing*, Chinese text online tokenisation platform and Python API, https://hanlp.hankcs.com/en/demos/tok.html

Hubert L. et Arabie P. (1985) Comparing partitions, *J. Classification,* 2, 193–218.

Kim J. et Park H., Toward Faster (2008) Nonnegative Matrix Factorization: A New Algorithm and Comparisons, *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, 353-362.

Lebart L. et Salem A. (1994). Statistique textuelle. Paris : Dunod.

Lee D. D. et Seung H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788-791.

Lelu A. (2008) La méthode de classification non-supervisée K-means axiales. *Rapport Technique*, 12 p. ⟨inria-00333865⟩

Lelu A. et Cadot M. (2009) Graphes des liens et anti-liens statistiquement valides entre les mots d'un corpus textuel. In *EGC 2009*, Strasbourg, 367-378. ⟨inria-00342751⟩

Lelu A. et Cadot M., (2011). Espace intrinsèque d'un graphe et recherche de communautés. In *Revue I3 - Information Interaction Intelligence*, 2011 (1), 1-25. ⟨hal-00641128⟩

Lelu A. et Cadot M. (2019) Evaluation of text clustering methods and their dataspace embeddings: an exploration. In *IFCS 2019 - 16th International of the Federation of Classification Societies*, Thessaloniki, Greece. ⟨hal-02116493v4⟩

Lelu A. et Roussanaly A. (2014) Espaces intrinsèques des relations entre mots : une exploration multi-échelle. In *JADT 2014 : 12e Journées internationales d'Analyse statistique des Données Textuelles*, Paris, 409-420. ⟨hal-01067984⟩

Lelu A., Zitt M. et Bassecoulard E. (2013) Robustesse des classements bibliométriques, à travers la convergence des thèmes obtenus par citations et lexiques. In *VSST 2013*, Nancy.

Maspéro G. (1879-86) Romans et poésies du papyrus Harris N° 500, In *Études égyptiennes*, Tome 1, 3 fascicules. Paris: Maisonneuve; Imprimerie nationale, 311 p.

Poe W. C. (2008) The Writing of a Skillful Scribe - An introduction to hieratic Middle Egyptian through the text of The Shipwrecked Sailor, Santa Rosa, 341 p., www.egyptologyforum.org/bbs/Stableford/

Taladoire E. (2012) Les trois codex mayas, Paris : Balland, 240 p.

Vanni L., Corneli M., Mayaffre D. et Precioso F. (2023) From text saliency to linguistic objects: learning linguistic interpretable markers with a multi-channels convolutional architecture. In *Corpus*, 24, ⟨10.4000/corpus.7667⟩. ⟨hal-04004208⟩