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 n-grammes, et de représentations multidimensionnelles graduées

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

[1] Université Grenoble Alpes (UFR de Langues étrangères) – brunodelprat@club-internet.fr

[2] Laboratoire LORIA, Nancy – martine.cadot@loria.fr

[3] Retraité, Université de Franche-Comté – alelu@orange.fr

Abstract

For ancient logosyllabic scripts, without separators between lexical units, 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 with mayaTeX, and their processing by Correspondence Analysis and graded unsupervised classification (Axial K-Means and Non-negative Matrix Factorization). We identify intra- and inter-text features 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, non-negative matrix factorization, NMF, intrinsic dimension, Monte-Carlo simulations, Tournebool algorithm.

Résumé

Pour des écritures logosyllabiques anciennes, ignorant les séparateurs entre unités lexicales, 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 avec mayaTeX, et leur traitement par Analyse Factorielle des Correspondances et classification non supervisée graduée (K-Moyennes Axiales et Non-negative Matrix Factorization). Nous en dégageons des manifestations, intra- et inter-textes, 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, 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, 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 « Livres des morts » sont les plus longs, comme le *papyrus de Nouou* (Budge, 1899) qui rassemble environ 93 000 hiéroglyphes en 131 chapitres, accompagnant un riche défunt avec les instructions à accomplir face au dieux lors de sa navigation éternelle. Les grands romans ou épopées sont absents ; les récits historiques ou littéraires sont courts, autour de 5000 hiéroglyphes, comme celui de notre corpus d'étude. Pour son déchiffrement au début du 19e siècle, Champollion s'est appuyé sur les textes de stèles, à quelque 1000 hiéroglyphes. Le cerveau humain déduit rapidement les règles du langage d'un nombre limité de situations rencontrées, sans recours à de grands volumes de parole ou de textes.

L'écriture logosyllabique maya est utilisée en de riches inscriptions sur plus de 13 siècles qui constituent néanmoins un volume plus faible de textes disponibles : trois manuscrits divinatoires (environ 24.600 glyphes), quelques milliers d'inscriptions courtes en caractères cursifs sur poteries et peintures murales, 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. Mais les besoins d'avancées sur les « Small Data » sont toujours prégnants, car dans certains secteurs, comme l'étude des textes antiques, on ne dispose de rien d'autre, et 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 : on peut envisager des milliers, voire dizaines de milliers de passages nous rapprochant de l'*optimum optimorum*. Créer des centaines ou milliers de clones aléatoires d'un tableau de données peut nous permettre de trouver le nombre maximum de classes pertinentes.

2. Spécificités des écritures logo-syllabiques et choix effectués

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 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, dotée 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'Égypte antique, des dictionnaires groupaient les hiéroglyphes par familles de représentations graphiques (Ciel, terre et eau ; l'Homme et ses

occupations ; les Oiseaux ;...). Le *Manuel de codage* (Buurman et al., 1988) a normalisé pour l'informatique un tel codage thématique, ici utilisé.

Le signe élémentaire maya est le glyphe, affixes plats qui s'organisent en tournant et par symétries autour d'éléments centraux presque carrés, le plus souvent des logogrammes correspondant à un morphème ou mot, servant souvent par effet « rébus » d'élément syllabique. Les affixes ont avant tout une valeur phonétique syllabique. Un ou plusieurs glyphes assemblés ensemble remplissent harmonieusement l'espace rectangulaire d'un cartouche. Les manuscrits mayas sont organisés en blocs de 2 à 12 cartouches qui constituent autant de phrases. Selon la place dans la page, le scribe entassait ou étalait 1 à 8 éléments de base par cartouche, groupés par expressions signifiantes, ou pour une belle mise en page. Le découpage en cartouche est une bonne approximation à celui des mots mayas, incluant préfixes ou post-fixes grammaticaux. Les textes mayas présentent un état condensé de la langue, sans flexions et particules grammaticales non indispensables, tout comme le chinois classique. Des chercheurs soviétiques (Évréïnov et al., 1961-69) ont codé informatiquement sur 3 chiffres les signes des manuscrits qui comprennent environ 520 glyphes, codage repris ici pour notre corpus maya.

2.2. Prétraitements des textes et visualisation des résultats hiéroglyphiques sous LaTeX

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, des glyphes fréquents présentent des variantes graphiques interchangeables, dont 98 assimilées dans notre catalogue des signes. 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 les traite souvent graphiquement comme un signe. Nous les avons décomposés en leurs éléments logiques dans notre corpus, ce qui réduit encore de 90 les glyphes élémentaires différents considérés 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. De façon similaire, le code vide "777" sert de 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.

Les paléographies hiéroglyphiques informatiques de nos deux corpus égyptien et maya sous LaTeX ont été constituées par rapport aux reproductions des manuscrits originaux , à l'aide du système mayaTeX (Delprat et Orevkov, 2007), outil original développé sous TeX de saisie et édition de textes hiéroglyphiques, composant les signes dans le cartouche maya tout comme le quadrat égyptien. Deux principaux opérateurs de composition des signes "." et ":" juxtaposent deux signes latéralement et verticalement, ou des groupes de signes à l'aide de parenthèses. Des polices de caractères mayas ont été élaborées, ici la police CODEX, et plusieurs polices égyptiennes, ici : IMPRIMERIE NATIONALE de style silhouette des inscriptions monumentales.

2.3. 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 : la phrase (égyptien) ou bloc (maya), et la rubrique (égyptien) ou almanach (maya), moins fin mais respectueux des notations des scribes. 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 méthodes KMA et NMF décrites plus loin, 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).

3. Principes des traitements

L'Analyse Factorielle des Correspondances est utilisée continuement depuis des décennies en analyse des données textuelles (Lebart et Salem, 1994). Essentielle, l'équivalence distributionnelle - fusionner 2 lignes ou colonnes de même profil relatif ne change pas l'analyse - est partagée par l'Analyse Factorielle Sphérique (AFS) (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, 1994), utilisent cette représentation et discriminent 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. Elle mesure la similarité du tableau reconstitué à celui d'origine, minimisant leur écart quadratique : en AFS et KMA la part de la somme des valeurs du tableau d'origine reconstituée.

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. En bref : la décomposition aux valeurs singulières, sous-jacente à l'Analyse en Composantes Principales (ACP) et à l'AFC, maximise la précision de reconstitution ; la NMF, dénuée de la contrainte d'orthogonalité 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 minimise comme les KMA l'écart quadratique avec le tableau d'origine. Nous comparerons plus loin sur un exemple ces deux représentations très proches : elles présentent des axes de classes gradués, et non des appartenances en tout ou rien comme les méthodes de clustering classiques.

Combien de classes pertinentes peut-on tirer d'un tableau de donnée ? Notre procédure TourneBool (Lelu et Cadot, 2011) permet de fixer de façon rigoureuse un nombre de classes pertinentes. Si ces classes regroupent des profils relatifs, on aura autant de dimensions distinctes de l'espace que de classes. Appelons dimension intrinsèque le nombre maximal de ces dimensions, nécessairement inférieur au minimum des nombres de lignes et de colonnes. Cadot (2005) a développé une méthode de Monte Carlo produisant à partir d'un tableau binaire (en pratique, l'aspect très clairsemé d'un tableau de données textuelles peut l'assimiler à sa contrepartie binaire), 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,

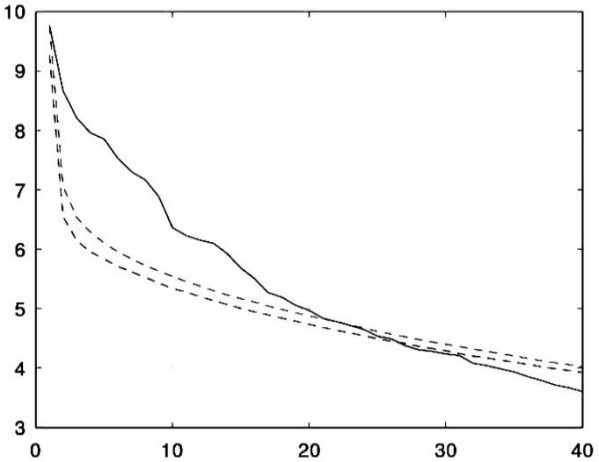

Figure 1 : Eboulis des valeurs propres du tableau 577 trigrammes × 143 phrases. En pointillés : intervalle de confiance à 95 % établi par simulations Monte Carlo. Conte du Naufragé.

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 fixe le nombre maximal de classes, i.e. d'axes ou facteurs, orthogonaux ou obliques, qu'on peut en extraire, ici 21 pour le classement des trigrammes. D'abord, fondées sur le nombre de phrases des corpus, six classes ont été choisies pour l'égyptien (seulement 143 phrases) ; et 15 classes dans le cas des 3 codex mayas (844 blocs-phrases). Puis, nous avons testé 21 classes pour le corpus égyptien, les résultats confirmant un compromis raisonnable entre finesse d'analyse et charge mentale d'interprétation, excessive au-delà de ~50.

4. Application des méthodes AFC et KMA à des corpus anciens

4.1. Le Conte égyptien du naufragé

Le papyrus Ermitage N° 1115 du Moyen empire (Golénischeff,1913), bien conservé et complet en écriture hiératique, comprend la seule version retrouvée de ce conte du début de la XIIe dynastie au 20e siècle av. J.-C. où les expéditions minières renforcent l'Égypte sur ses voisins. Les phrases sont écrites par le scribe toutes à la suite sans respirations, exceptés 20 segments marqués par les rubriques à l'encre rouge de leurs premiers mots, pour 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), les égyptiens ont marqué de points rouges la fin des mots au dessus du texte, comme aide à la lecture.

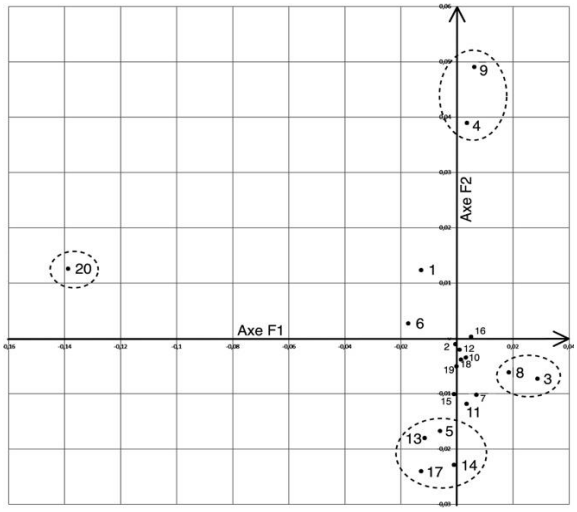

Figure 2 : Diagramme des axes F1 et F2 par analyse AFC des rubriques du corpus égyptien

Les groupements par analyse AFC selon l'axe F2 des rubriques (Figure 2) permet de dégager les associations suivantes : énumérations de produits et animaux (Rub. 17, 14, 13 et 5, les autres rubriques n'en contiennent pas) ; la tempête et en conséquence le naufrage (Rub. 4 et 9, on n'en parle pas ailleurs) ; le bateau et son équipage (Rub. 3 et 8, on n'en parle pas ailleurs) ; genre de notice bibliographique (Rub 20), écrite dans un style différent de celui du récit. Dans un texte littéraire, 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, c'est la mise en abyme du récit du serpent roi/dieu imbriqué à celui du naufragé.

L'étude rhétorique et quantitative des tutoriels en réalité virtuelle par AFC (Dozo et Barnabé, 2022) montre que le « hub », sorte d'écran ou plateforme de navigation de ces jeux, instaure

une mise en abyme, comme médiation régulièrement employée pour apprendre au joueur à « prendre » le jeu en mains. 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 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 largement informative.

Une analyse par KMA (optimisation sur 1000 passages) sur les 577 trigrammes (pleins) de signes hiéroglyphiques croisés avec les 143 phrases, sans découpage en mots logiques a été réalisée, avec appartenance d'un trigramme à plusieurs classes si son indice de centralité (Lelu et al., 2013) y est > 0,15. Il en résulte que 27 trigrammes, soit 5 % sont attribués au moins à 2 classes. Ces trigrammes multiclasses sont moins nombreux qu'anticipé. En corpus égyptien, l'approche multiclasse dégage mieux les expressions de plusieurs mots formant les têtes de classes, par rapport à l'attribution des trigrammes à une classe unique. Trait rassurant, la forte contribution du trigramme phonétique M17.M17.X1 à quatre 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 à partir d'un trait répétitif de morphologie interne.

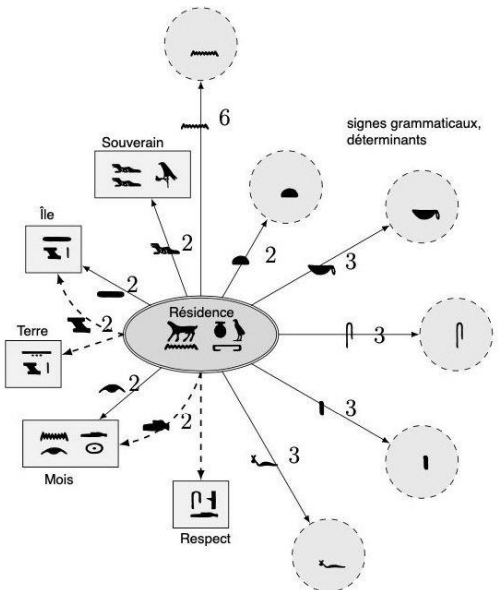

Figure 3 : Voisinage par AFC du terme « résidence ». Conte égyptien du Naufragé.

Dans l'approche découpage en phrases sans séparateurs de mots logiques pour le 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.

Avec une analyse AFC sur des quadrigrammes du type : "X_Y (2 à 5 trous) P_Q", soit un bigramme plein suivi de 2 à 5 "trous" et d'un autre bigramme plein, 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 sont en pointillés ; au centre le terme « résidence ». Le 2^{e} bigramme plein 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. Dix 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 ces 4-grammes 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 sur phrases logiques et 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 centralités 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 ».

Indice de centralité	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 sur corpus égyptien.

Une seule classe de phrases est caractéristique de sa structure avec une centralité 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 centralité des trigrammes > 0,15 à 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. Les axes thématiques issus du passage KMA sur 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 3-grammes à fonction grammaticale avec centralité < 0,6 dans les clusters, alors qu'ils peuplaient avec les 3-grammes « à cheval » 60 % des classes dans les passages sans séparation des mots.

4.2. Les 3 codex Mayas

Notre corpus comprend les trois seuls manuscrits mayas (Taladoire, 2012) sauvés des *autodafé* des évangélisateurs ou du climat humide d'Amérique centrale, préservés dans les bibliothèques de villes d'Europe. Ce sont les codex de : Paris (13^{e} siècle), Dresde (15e siècle) et Madrid (16^{e} ou 17^{e} siècle, partiellement saisi dans notre corpus). Sous forme de 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 compte 957 blocs phrases, sur 137 almanachs.

Indice de centralité	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 sur corpus maya.

Nos premiers traitements KMA sur le *codex de Dresde* (Hallab et al., 2010) étaient fondés sur un découpage des folios en registres supérieur, médian et inférieur. Ici, ils sont regroupés en almanachs complets sur lesquels est effectuée une analyse KMA (optimisation sur 1000 passages) en 15 classes de bigrammes (pleins) de glyphes mayas en fonction des 137 almanachs et des cartouches de texte, sans tenir compte du découpage en blocs-phrases, ni des bigrammes à cheval sur plusieurs cartouches. L'analyse obtenue sur les 810 bigrammes dégage des axes thématiques : Axe 1 « résonna et augures » (s'étend sur 60 almanachs), Axe 9 « les dignitaires et *Kauil*, dieu de la fertilité » (Figure 5, s'étend sur 74 almanachs). Les dates et nombres distances ont ici été expurgés du corpus. Un premier passage KMA les associant aux textes des almanachs avait produit des axes dont certains ne contenaient que ces nombres et dates, fortement présents dans les autres axes aussi, devenant difficilement interprétables.

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

Dans (Lelu et Cadot, 2021) 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, comme 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 du tableau binarisé (il ne comporte que 49 valeurs autres que un, sur 1756 autres que zéro). Celles-ci partagent la même répartition de densité que le tableau de référence : 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. C'est pour la 21^{e} valeur propre (par ordre décroissant) que l'« éboulis » des valeurs propres de la matrice binarisée d'origine (Figure 1) atteint ce couloir. Comparons maintenant les deux méthodes en deux temps.

D'abord en leur imposant 6 axes ou classes, comme précédemment : on applique les KMA avec 1000 passages et des graines d'initialisation de 1 à 1000. Le meilleur passage obtient un taux de reconstitution des données de 416,5/1756=23,72 %. Pour l'équité du test, on transforme ensuite les données pour obtenir avec les NMF (Kim et Park, 2008) une même fonction objectif, ici le pourcentage de la somme des données reconstituées : $X \rightarrow X^{(1/2)}$, pour 1000 passages également. Le taux de reconstitution est de 23,79 %, 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. Leurs 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.

Dans un deuxième temps, en se plaçant dans l'espace intrinsèque des données (soit avec 21 axes/classes), les KMA aboutissent à un taux de 47,61 % de reconstitution des données, contre 50,70 % avec la factorisation NMF dans l'espace des données transformées comme ci-dessus, avec NMI=0,7992, ARI=0,6279, et ISAR=0,5933 (obtenu sur 1000 passages). D'où l'hypothèse qu'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, optimisées avec 200 passages, les taux de reconstitution sont comparables – respectivement 47,36 % et 50,47 %, mais les KMA prennent sur un même ordinateur (portable Dell Latitude E4300) 7,3 sec. contre 105,3 sec., soit 36 fois moins de temps de calcul.

6. Conclusions

Sur des données de taille limitée par nature, issues du codage et pré-traitement de deux corpus en écritures logosyllabiques, l'égyptien et le maya, nous avons appliqué l'AFC, puis deux algorithmes de classification non-supervisée, les KMA et la NMF, aboutissant au même type de représentation nuancée, qui ajoutent à la création de classes leur représentation de type factorielle. En ignorant la frontière des mots égyptiens ou des cartouches mayas, aussi bien l'approche AFC que KMA révèle principalement la structure du récit avec un découpage en rubriques ou almanachs, ou rapproche des expressions similaires sur un découpage en phrases. En séparant les mots ou cartouches, le champ sémantique est privilégié avec des classes thématiques comportant moins de 10 % d'éléments grammaticaux et dont les n-grammes avec indice de centralité < 0,15 environ constituent du bruit en fonds de classes. Séparer ou pas les 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 de la NMF dépassent de peu ceux des KMA, alors que cette dernière méthode est bien moins gourmande en temps de calcul, y compris quand il s'agit d'extraire le maximum d'information pertinente possible, quantifiée par la notion de dimension intrinsèque d'un tableau de données binaires. Le processus d'extraction de cette dimension et l'optimisation poussée avec les méthodes itératives utilisées profitent du caractère « Small Data », inenvisageable à l'échelle des « Big Data ». Nous avons mis en ligne sur HAL (https://hal.science/hal-04523153) sous licence GPL les codes Python et Octave utilisés, ainsi qu'un exemple de chaîne de traitements. Pour résumer, on peut dire que les KMA appliquées à un tableau X constituent une version approchée de la NMF du tableau $X^{(1/2)}$ avec plus de clarté : indices de centralité des lignes et des colonnes homogènes, aux valeurs entre 0 et 1, et avec une fonction objectif simple (i.e. taux de reconstitution des données).

Un réglage différencié des paramètres d'exécution adapte de façon empirique par tâtonnements nos méthodes aux spécificités de différentes écritures logosyllabiques, comme les n-grammes « à trous » pour l'exploration par AFC de quelques voisinages linguistiques, pour systématiser les contextes structuraux et stylistiques. 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, et de même par les KMA ou NMF. Le travail sur les trois codex mayas réunis pour la première fois sous une forme homogène et exploitable par ordinateur ne fait que commencer.

Bibliographie

Budge E. A. W. (1899). *The Book of the Dead: Facsimiles of the Papyri of Hunefer, Anhai, Kerasher and Netchemet, with supplementary Text from the Papyrus of Nu*. Londres : British museum.

Buurman J., Grimal N., Hainsworth M., Hallof J. et van der Plas D. (1988). *Inventaire des signes hiéroglyphiques en vue de leur saisie informatique*, Mémoires de l'Académie des Inscriptions et Belles Lettres (nouvelle série), 8. Paris : Institut de France.

Cadot M. (2005). A Randomization Test for extracting Robust Association Rules. In *3rd World Conference on Computational Statistics & Data Analysis (CSDA 2005)*. Limassol. inria-00337069.

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

Danon L., Díaz-Guilera A., Duch J. et Arenas A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics*, 9, 219-228.

Delprat B. et Orevkov S. (2012). MayaPS: Typing Maya Hieroglyphics with TeX/LaTeX. *TUGboat*, 33 (3), 289-294.

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. https://doi.org/10.4000/sdj.4098.

Évréïnov E. V., Kosarev Y. G. et Oustinov V. A. (1961-69). Применение электронных вычислительных машин в исследовании письменности древних майя, Новосибирск: АН СССР [Utilisation des ordinateurs pour les recherches sur l'écriture des anciens Mayas, 4 vol.]

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

Greenacre M et Hastie T. (1987). The geometric interpretation of correspondence analysis. *Journal of the American Statistical Association*, 82 (398), 437-447.

Hallab M., Delprat B. et Lelu A. (2010). Codage et classification non supervisée d'un corpus maya. In Ben Yahia S., Petit J.-M., Collectif Cépaduès (Eds.), *Extraction et gestion des connaissances (EGC 2010)*, Revue des Nouvelles Technologies de l'Information, RNTI-E-19, 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. (2008). Toward Faster Nonnegative Matrix Factorization: A New Algorithm and Comparisons, In Giannotti F et al.. (Eds.), *Proc. 2008 Eighth IEEE Int. Conf. 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. (1994). Clusters *and* factors: neural algorithms for a novel representation of huge data sets. In Diday E. (Ed.), *New approaches in classification and data analysis*. Berlin : Springer, 241-248.

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

Lelu A. et Cadot M. (2021). Evaluation of text clustering methods and their dataspace embeddings, In Chadjipadelis T. et al. (Eds.), *IFCS 2019: Data Analysis and Rationality in a Complex World*, 131-139.

Lelu A. et Roussanaly A. (2014). Espaces intrinsèques des relations entre mots : une exploration multi-échelle. In Née E et al. (Eds.), JADT 2014. 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. *VSST 2013*, Nancy. hal-00926631.

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 : W. Poe, www.egyptologyforum.org/bbs/Stableford/.

Taladoire E. (2012). *Les trois codex mayas*. Paris : Balland.

Vanni L., Corneli M., Mayaffre D. et Precioso F. (2023). From text saliency to linguistic objects. *Corpus*, 24, [en ligne] 10.4000/corpus.7667. hal-04004208.