Détection automatique des textes épistolaires du corpus néo-égyptien : méthodes exploitant la récurrence de motifs discriminants

Stéphanie Gohy[1], Benjamin Martin Leon[2]

[1]Aspirante FNRS, Université de Liège – sgohy@ulg.ac.be
[2]Université de Liège – b.martinleon@ulg.ac.be

Abstract

In this paper, we will develop two methods allowing an automatic detection of the Late-Egyptian epistolary genre. Among the criteria which could be mobilized to identify different genres within a corpus, the study of "motifs" ("patterns") represents a particularly promising approach that has already been successfully exploited for a corpus of Latin texts. In our communication, we suggest applying this process to the Late-Egyptian corpus, and more particularly to the epistolary genre. Two methods will be applied to our corpus to identify whether or not particular documents belong to the epistolary genre. We shall begin by explaining the principle of functioning of these two methods. The results obtained will then be analyzed; we shall try to understand why certain documents were improperly classified.

Résumé

Dans cette contribution, nous proposons de développer deux méthodes permettant une détection automatique du genre épistolaire néo-égyptien. Parmi les critères pouvant être mobilisés afin de rendre compte des différents genres composant un corpus figure notamment l'étude de « motifs », procédé déjà appliqué, entre autres, à un corpus d'historiens latins[1]. Dans notre communication, nous proposons d'appliquer ce procédé au corpus néo-égyptien, et plus particulièrement au genre épistolaire. Pour ce faire, deux méthodes seront appliquées à notre corpus afin d'identifier les documents appartenant ou non au genre épistolaire. Nous commencerons par expliquer le principe de fonctionnement de ces deux méthodes. Les résultats obtenus seront ensuite analysés ; on tentera de comprendre pourquoi certains documents ont été mal classés.

Mots-clés : motif, détection automatique, classification, genre épistolaire, néo-égyptien, arbre de décision

1. Introduction

Cette contribution a pour objectif la description de deux méthodes visant à l'identification de motifs discriminants dans le corpus annoté Ramsès ; elle s'intègre dans le cadre plus général d'une étude s'attelant à la catégorisation automatique du corpus néo-égyptien dans son ensemble.

On présentera dans un premier temps le corpus qui est en cours d'annotation dans le cadre du projet Ramsès et l'état de langue concerné (§2). Ensuite, si l'on sait que, dans tout processus

1 Voir les diverses études réalisées par Longrée D., Mellet S., Luong X. et Barthélémy J.-P. (cf. références).

de catégorisation d'un corpus, plusieurs critères peuvent être mobilisés afin de rendre compte des diverses formes textuelles attestées (structure formelle, contexte d'énonciation, répertoire, phraséologie, etc.), c'est l'étude des motifs, compris comme l'*association récurrente de n éléments de l'ensemble E*[2] qui sera ici envisagée dans cette optique. Au cœur de la présente contribution, un motif est plus spécifiquement défini comme étant une séquence linéaire de lemmes. Cependant, un motif pourrait-être défini selon plusieurs dimensions, en faisant varier la nature (lemme, partie du discours, ...) des éléments qui le composent. Toutefois, comme précisé dans notre définition ci-dessus, nous n'envisageons pas ici de motifs multidimensionnels. Pour chacune des deux méthodes qui vont être exposées (§3 et §4), il s'agira dans un premier temps d'effectuer une détection automatique des motifs caractérisant le genre épistolaire. Cette opération a pour but d'identifier, sans définition préalable, les motifs récurrents qui sont caractéristiques des lettres, mais également ceux qui sont caractéristiques des textes qui ne sont pas des lettres. Il s'agira ensuite, dans un second temps, d'opérer une classification automatique des textes qui n'ont pas été utilisés dans la première étape afin d'établir l'appartenance ou non d'un texte au genre épistolaire, et cela, sur base des motifs extraits précédemment. Cette étape permettra d'évaluer la pertinence des motifs extraits automatiquement dans une perspective classificatoire.

2. Ramsès : un corpus annoté du néo-égyptien

Depuis quelques années (i.e. fin 2006), le service d'égyptologie de l'université de Liège concentre ses travaux sur un projet intitulé « projet Ramsès »[3]. L'objectif est de développer un corpus richement annoté du néo-égyptien :

- Les textes du corpus sont encodés sous leur forme hiéroglyphique et enrichis d'annotations ecdotiques et linguistiques multiples (analyse morphologique détaillée et lemmatisation) ainsi que de traductions alignées au niveau propositionnel et d'informations bibliographiques permettant de donner une assise aux choix philologiques que les encodeurs sont amenés à prendre. Le corpus devrait, à terme, rassembler la totalité des textes conservés en néo-égyptien ; aujourd'hui, il comprend environ 1300 textes ; 8000 lemmes ; 400000 mots-occurrences, ce qui représente environ 35% de l'ensemble du matériau linguistique conservé pour la période.
- Le néo-égyptien est la langue conservée dans la majorité des sources textuelles provenant d'Égypte ancienne et datant du Nouvel Empire et de la Troisième Période Intermédiaire (XVIII^{e}-XXV^{e} dynastie ; *c.* 1500-700 av. J.-Chr.). Le corpus néo-égyptien est constitué de plusieurs milliers de documents écrits dans les registres les plus divers : on dénombre des œuvres littéraires, des textes hymniques et eulogiques, des textes administratifs, juridiques, économiques, des lettres, des textes scolaires, etc.

Parmi les différentes formes textuelles repérées dans le corpus, cette contribution se concentrera essentiellement sur la forme épistolaire. Deux raisons ont présidé au choix de ce sous-corpus particulier ; tout d'abord, c'est l'un des corpus les plus complets actuellement encodé dans la base Ramsès (plus de 95% de textes conservés sont inclus dans le corpus). Par ailleurs, il s'agit

2 Longrée D., Luong X., Mellet S., 2008.

3 Rosmorduc S., Polis St., Winand J., 2009 ; Winand J., Polis St., Rosmorduc S., sous-presse ; Polis St., Winand J., Honnay A.-C., 2010.

d'un corpus relativement étendu (474 textes) et il se prête donc particulièrement bien à l'étude de motifs récurrents et discriminants.

3. Première méthode : Classification des textes sur base d'une liste exhaustive de motifs combinée à une heuristique

3.1. Principe

La première méthode se déroule en deux phases. La première phase consiste à générer une liste exhaustive de toutes les sous-séquences de lemmes contenues dans 70% des textes de la base (lettres et non lettres confondus). On obtient alors une liste de motifs (séquences de lemmes) qui sont présents uniquement dans les lettres, uniquement dans les non lettres ou dans ces deux types de textes à la fois. La seconde phase consiste à utiliser cette liste de motifs pour classer les 30% de textes restant. Ce classement est effectué au moyen d'une heuristique.

3.2. Génération de la liste exhaustive des motifs

Une méthode efficace a dû être déployée afin d'extraire toutes les sous-séquences de lemmes des textes. Cette méthode devait également associer à chacune des sous-séquences extraites l'information de la proportion de textes qui sont des lettres et de textes qui ne sont pas des lettres qui la contiennent. La méthode qui a été retenue est la construction progressive d'un arbre des suffixes adapté à nos besoins.

3.2.1. Construction de l'arbre des suffixes

L'arbre est construit de la façon suivante. Chacun des textes faisant partie de l'ensemble d'apprentissage (70% des textes de la base) est parcouru proposition par proposition. Chacun des suffixes de ces propositions est inséré dans l'arbre comme suit. En partant de la racine de l'arbre, le premier lemme du suffixe courant est associé à une branche partant de la racine. Si une telle branche n'existe pas, alors elle est créée. Le deuxième lemme du suffixe courant est associé à une branche dont l'origine est le nœud d'arrivée de la branche associée au lemme précédent (dans ce cas-ci, le premier lemme du suffixe courant). Les autres lemmes du suffixe courant sont ajoutés de la même manière, en descendant en profondeur dans l'arbre, et en créant les branches nécessaires. En visitant chacun des nœuds de l'arbre (pendant cette même construction des branches successives de l'arbre), l'identifiant du texte qui contient la séquence de lemmes correspondant au nœud visité (c'est-à-dire la séquence de lemmes obtenue en parcourant l'arbre de la racine à ce nœud) est associé à ce nœud, si ce n'est déjà fait. Sur base de ces identifiants conservés dans les nœuds de l'arbre, on connaît donc pour chacune des séquences de lemmes allant de la racine à un nœud quelconque de cet arbre, le nombre de textes qui sont des lettres et le nombre de textes qui ne le sont pas, qui contiennent la séquence en question. La liste des motifs est alors obtenue en parcourant l'arbre de la racine à chacun des nœuds de ce dernier. En effet, étant donné que les séquences de lemmes définies par la suite des branches de l'arbre allant de la racine à chacune des feuilles de ce dernier correspondent à tous les suffixes des propositions des textes, en prenant les préfixes de ces suffixes on arrive à extraire toutes les sous-séquences de lemmes effectivement contenues dans les textes.

3.2.2. Structure de la liste des motifs

La liste de motifs ainsi obtenue est évidemment conséquente, et la grande majorité des motifs contenus dans cette liste ne sont pas pertinents pour déterminer si un texte appartient ou non au genre des lettres. Pour identifier les motifs marqueurs du genre des lettres (ou du genre des non lettres), on va utiliser les informations suivantes pour un motif donné :

- la proportion de textes *lettres* qui contiennent ce motif : p(L) ;
- la proportion de textes *non lettres* qui contiennent ce motif : p(N) ;
- le rapport entre ces deux grandeurs : p(L) / p(N) = R.

Les différentes valeurs pouvant être prises par R sont les suivantes :

- **l'infini** (p(N) = 0)
  → le motif associé est présent uniquement dans les lettres ;
- **0** (p(L) = 0)
  → le motif associé est présent uniquement dans les non lettres ;
- **> 1**
  → le motif associé est davantage présent dans les lettres ;
- **< 1**
  → le motif associé est davantage présent dans les non lettres.

En triant la liste par ordre de valeurs de R décroissantes et valeurs de p(L) décroissantes (resp. p(N) croissantes) lorsque R > 1 (resp. R < 1), on obtient une liste ordonnée de motifs dont les premiers sont les plus marqueurs des lettres et dont les derniers sont les plus marqueurs des non lettres.

3.3. Test des motifs

Afin de tester la pertinence des motifs considérés comme étant les plus discriminants, on effectue un classement des 30% de textes restant (ensemble de test) en utilisant ces motifs via une heuristique.

3.3.1. Classification d'un texte

Les étapes pour attribuer un genre (lettre ou non lettre) à un texte T sont les suivantes. On commence par extraire tous les motifs de T (toutes les sous-séquences de lemmes contenues dans les propositions de T). On confronte ensuite chacun des motifs extraits de T aux motifs de la liste précédemment construite. Pour un motif de T, si ce motif est présent dans la liste, alors on incrémente les deux valeurs *valeur lettre* et *valeur non lettre* en fonction des valeurs associées à ce motif dans la liste, et cela en se basant sur une heuristique décrite dans la section suivante. Lorsque tous les motifs de T ont été confrontés à notre liste de motifs, on attribue à T le genre (lettre ou non lettre) correspondant à la valeur la plus élevée entre *valeur lettre* et *valeur non lettre*.

3.3.2. Heuristique et performance

L'heuristique qui a été retenue est très probablement perfectible mais donne déjà des résultats encourageants. Cette heuristique est la suivante :

1. rac = 4 ;
2. valeur lettre = valeur lettre + ;
3. si p(L) > 0.01 et p(N) < 0.005 → rac = 1 ;
4. valeur non lettre = valeur non lettre + .

Cette heuristique attribue un poids plus important à la proportion de lettres p(L) si cette dernière est significativement plus élevée que la proportion de non lettres p(N).

La performance obtenue avec cette heuristique est plutôt satisfaisante puisqu'elle permet de classer environ 90% des textes de l'ensemble de test de manière correcte.

3.4. Analyse des données

Le corpus épistolaire néo-égyptien est caractérisé par de nombreuses formulations récurrentes ; la phraséologie constitue donc un bon critère d'identification du genre, ainsi que l'illustre les résultats obtenus par cette première méthode. La formulation la mieux identifiée est l'expression wnn tAy.k Sa.t (Hr) spr r.k iw.k (Hr)… « Dès que ma lettre te parviendra, tu feras... ». Attestée à de nombreuses reprises, cette formulation caractéristique est employée dès la 19e dynastie ; on la rencontre principalement dans les lettres sur papyri des 20e et 21e dynasties. Parmi les formulations identifiées par cette première méthode, on reconnaît encore une tournure caractéristique employée dans la formule finale du genre épistolaire, nfr snb.k « Puisse ta santé être bonne », formule la plus régulièrement employée durant l'époque ramesside. À dire vrai, les lettres présentent une structure formelle assez stable : formule introductive (incipit) – salutations – corps de la lettre – formule finale – adresse. Il s'agit, ici, de la structure type ; bien évidemment, chacune de ces parties n'est pas systématiquement présente. En réalité, peu de lettres comportent ces 5 parties et l'absence de certaines d'entre elles s'explique par divers facteurs[4] :

- le support employé (papyrus >< ostracon) ; place considérablement réduite sur ostracon peut engendrer restriction à certaines parties ;
- le genre ; dans le corpus épistolaire, on relève 4 genres entre lesquels peuvent apparaître des distinctions : lettres – lettres-modèles – lettres littéraires – lettres aux morts ;
- l'époque de rédaction.

Les expressions couramment employées dans le registre des salutations sont également identifiées, telle que cette formule caractéristique des lettres de la fin de la 20e dynastie et de la 21e dynastie, twi (Hr) Dd n Dieu(x) imy n.k anx wDA snb aHaw qA iAw.t aA.t nfr Hs.t qnw aSA.t m-bAH + Dieu « Je dis à Dieu(x) qu'il(s) t'accordent vie, santé et prospérité, un temps de vie élevé, une longue et belle vieillesse, de nombreuses louanges en présence de

4 Bakir A., 1970, p. 31 : *These parts may not all occur in any one letter. Their presence and length are determined largely by the relations existing between the correspondents, the material on which the missive is written, as well as the character and purpose of the letter*

Dieu », ou cette expression employée durant toute l'époque ramesside, twi Hr Dd n Dieu(x) imy snb.k imy anx.k … « Je dis à Dieu(x) qu'il(s) t'accorde(nt) la santé, qu'il(s) t'accorde(nt) la vie,… ». À l'inverse, cette méthode fait également apparaître des motifs caractéristiques d'autres genres, jamais attestés dans le corpus épistolaire. Ainsi, le groupe Hr nxt kA est plutôt typique des documents royaux ; il s'agit d'une désignation du roi. De même, le motif anx n nb est caractéristique du genre des serments.

Si cette première méthode identifie une grande majorité des lettres, certaines ne sont pas reconnues. Parmi les quelques lettres mal classées, la plupart concernent des cas particuliers (documents fragmentaires, lettre littéraire). Considérons les exemples qui suivent :

La lettre conservée sur T. Caire JE 92920 n'est pas identifiée ; si ce document est bien une lettre, il s'agit, d'un texte datant du Moyen Empire[5]. Si l'on regarde ce document de plus près, on voit apparaître de nombreux points de divergences avec les lettres du Nouvel Empire ; ce sont probablement ces écarts qui ont mené à la non reconnaissance de ce texte comme faisant partie du genre épistolaire. On relève ainsi les points de divergence suivants :

- Ligne 1 : bAk n pr D.t iy Dd « Le serviteur du domaine d'éternité, iy, dit » ; notre document s'ouvre donc par un incipit, comme c'est souvent le cas dans les lettres du Nouvel Empire. Diverses formules peuvent apparaître dans cette partie de la lettre ; parmi les différents motifs identifiés, quelques-uns correspondent, d'ailleurs, aux formulations employées en incipit. Notre tablette s'ouvre par un incipit de type NP (exp.) Dd ; le nom du destinataire n'est donc pas mentionné. Dans les lettres du Nouvel Empire, on relève plusieurs incipit recourant au verbe Dd « dire », toutefois, dans ces différents types d'incipit, le nom du destinataire est systématiquement précisé ; au Nouvel Empire, la formule se présente donc sous la forme suivante : NP (exp.) Dd n NP (dest.). Les quelques motifs identifiés susceptibles de correspondre à la formulation employée dans la tablette sont donc systématiquement plus long ; on trouve toujours le nom de l'expéditeur suivi du verbe Dd et de la préposition n « à » introduisant le nom du destinataire. Par ailleurs, on remarquera que les quelques motifs identifiés correspondant à la formule d'incipit du Nouvel Empire (NP Dd n NP) se révèlent, certes, discriminants6 mais semblent peu attestés7. En effet, dans les quelques motifs correspondant, chacun conservent le nom des protagonistes ; or, si le corpus épistolaire néo-égyptien rassemble un grand nombre de lettres, on relève peu de cas d'échanges répétés entre les mêmes individus8. En d'autres termes, à chacune des lettres s'ouvrant par cette formulation correspondent autant de motifs ; afin que ce motif soit véritablement pertinent dans l'identification du genre épistolaire, il faudrait donc nécessairement inclure dans notre processus des catégories sous spécifiées. Ainsi, on pourrait imaginer un motif mêlant lemmes et catégories sémantiques, du type : nom propre – Dd « dire » - nom propre ;

5 Ce texte fut encodé dans la base de données lors d'une phase de test. À terme, il sera retiré de la base Ramsès qui se limite aux documents néo-égyptiens.

6 Ils sont seulement attestés dans le genre épistolaire (R = infini).

7 Chacun des motifs correspondant n'est jamais attesté dans plus de 5 lettres sur 100 = $p(L)$.

8 À vrai dire, mis à part les nombreux échanges entre les scribes Djéhoutymose et Boutéhamon, nous ne connaissons pas d'autres cas.

- Ligne 2 : swDA-ib pw n nb anx wDA snb « C'est une information pour le maître, vie, santé, force ». Notre tablette emploie à plusieurs reprises cette formulation. Le lemme swDA-ib figure parmi les motifs identifiés. Ainsi, on le rencontre dans deux formulations caractéristiques des lettres du Nouvel Empire : il est employé dans l'incipit du type NP (exp.) Hr swDA-ib n NP (dest.) « NP informe NP » ou dans la formule ky swDA-ib n pAy.i nb « autre information pour mon maître ». Cette dernière formulation est probablement la variante récente de la tournure employée dans notre lettre du Moyen Empire. La formule employée dans la tablette présentant des variantes importantes par rapport aux tournures attestées au Nouvel Empire, identifiées comme motifs discriminants, n'a donc pas été reconnue ;
- Ligne 3 : hAw nb n nb anx wDA snb aD wDA « Toutes les affaires du maître, vie, santé, force, sont saines et florissantes ». Si la lettre de la tablette Caire recourt à un lexique caractéristique de l'époque de rédaction, certains lemmes ne sont plus vraiment en usage au Nouvel Empire. Ainsi, dans ce passage où l'expéditeur rassure son maître sur la bonne conduite de ses affaires, on trouve encore le verbe aD « être sain et sauf ». Dans les lettres du Nouvel Empire, on s'attendrait plutôt à trouver une formule recourant à l'adverbe m-sSr « excellemment », sur le modèle de la construction suivante : SN nty r-xt pAy.i nb m-sSr « SN qui est sous l'autorité de mon maître va bien ». Par ailleurs, le groupe r-xt pAy.i nb m-sSr figure parmi les motifs discriminants du genre épistolaire ; il est seulement attesté dans le genre des lettres (R=infini) et se rencontre dans une moyenne de 17 lettres sur 100 (p (L)=0.16997168). Notre lettre du Moyen Empire recourant à une phraséologie ancienne, aucun motif n'a pu être identifié dans cette proposition.
- Ligne 5 : m Hs.t n.t Xnmw « dans les louanges de Khnoum ». Cette formulation, caractéristique du registre des salutations, est d'un emploi très régulier dans les lettres du Nouvel Empire. Toutefois, à cette époque, on trouve plutôt m Hs.t n imn-ra « dans les louanges d'Amon-Rê »[9]. La mention du dieu Khnoum est exceptionnelle au Nouvel Empire ; par ailleurs, s'il apparaît quelquefois dans le registre des salutations, il est toujours accompagné d'autres divinités. Parmi les motifs identifiés, un aurait pu servir à identifier notre tablette. Ainsi, on relève le motif suivant m[dans, en tant que,de] - Hs.t[éloge, faveur] - n(j)[de] ; il s'agit donc exactement des premiers lemmes de notre formule. Ce motif est caractéristique des lettres (R=infini), toutefois, il n'a pas permis l'identification de la tablette. La méthode employée ne faisant pas apparaître distinctement les motifs retenus afin d'identifier un document comme appartenant ou non au genre épistolaire, on ignore, à vrai dire, si ce motif figurait parmi ceux jugés pertinents. En tous les cas, sous cette forme, le motif se rencontre à peine dans une moyenne de 8 lettres sur 100 (p (L) = 0.008498584).
- Ligne 11 : nfr sDm nb anx wDA snb « Puisse l'écoute du maître, vie, santé, force, être bonne ». La lettre de la tablette Caire s'achève donc par une formulation caractéristique de l'époque de rédaction. Les lettres du Nouvel Empire peuvent, éventuellement, comporter une formule finale ; à cette époque, la tournure la plus fréquente est nfr snb.k « Que ta santé soit bonne », forme récente de nfr sDm.k « Puisse ton écoute être bonne ». À la formule nfr snb.k correspond, par ailleurs, un motif discriminant ; ainsi, le motif

9 Un motif correspond, par ailleurs, à cette formule.

nfr[parfait, bon] - snb[santé] est limité au lettres (*R* = infini) et d'un emploi régulier (*p (L)* = 0.014164306). À l'inverse, aucun motif ne correspond à la formulation employée dans notre tablette ; à vrai dire, cette tournure ancienne se rencontre exceptionnellement au Nouvel Empire[10]. Par ailleurs, dans les quelques cas relevés, le sujet est toujours de nature pronominal ; on trouve donc plutôt la formulation suivante : nfr sDm.k « Puisse ton écoute être bonne » ;

À l'inverse, cette méthode a également identifié des textes comme faisant partie du genre épistolaire mais qui ne sont pas des lettres. À nouveau, il s'agit souvent de cas particuliers ; la plupart des documents erronément identifiés au genre épistolaire sont en fait des questions oraculaires, des textes extrêmement courts, se limitant parfois à un lemme, donc difficilement identifiables dans le cadre d'une reconnaissance automatique reposant sur l'occurrence de motifs particuliers. À titre d'exemple, comparons les deux passages suivants ; le premier est issu d'une question oraculaire erronément identifiée au genre épistolaire. Le second exemple est tiré d'une courte lettre se limitant à l'incipit (type n NP) et une question au destinataire concernant la valeur d'une natte :

Ex. : in pA dAiw (m)-di xa
« Est-ce que le pagne est en possession de xa ? » O. DeM 797, r°1
Ep. Ramesside – question oraculaire

Ex. : n nxt-imn
in p(A)y tmA m wa(.t) ip.t it
« À nxt-imn,
Est-ce que cette natte vaut un oipé de grains ? » O. DeM 784, r° 1
Ep. Ramsès II – lettre

Dans le premier exemple, erronément classé comme une lettre, un seul motif a été identifié par la méthode, il s'agit de la particule interrogative in. Toutefois, ce motif n'est pas discriminant puisqu'il est autant attesté dans les lettres que dans les autres genres (*p (L)* = 0.10764872 / *p (N)* = 0.13799623). Dans le second exemple, la présence de l'incipit nous permet d'affirmer qu'il s'agit bien d'une lettre ; sans cela, on pourrait toujours hésiter entre le genre épistolaire et le genre de la question oraculaire.

4. Seconde méthode : classification des textes au moyen d'un arbre de décision

4.1. Principe

La seconde méthode se déroule également en deux phases. La première phase (phase d'apprentissage) consiste à générer un arbre de décision modélisant le processus de classification d'un texte, et cela sur base d'un fichier de données. Ce fichier de données est constitué à partir de 70% des textes de la base et liste pour chacun d'entre eux les différents motifs qu'il contient et le genre qui lui est associé (lettre ou non lettre). La seconde phase (phase de test) consiste à classer les 30% de textes restant suivant le processus de décision imposé par l'arbre.

10 Dans le corpus épistolaire du Nouvel Empire, on relève à peine 4 cas.

4.2. Phase d'apprentissage

4.2.1. Description de l'arbre de décision

À chacun des nœuds de l'arbre est associé un ensemble de textes. À chacun des nœuds internes de l'arbre (tous les nœuds exceptés les feuilles) est associée une variable de segmentation qui va permettre de séparer les textes dans les différents nœuds fils de ce nœud. Les différentes branches partant d'un nœud vers ses fils correspondent aux différentes valeurs pouvant être prises par la variable de segmentation. À chacune des feuilles de l'arbre (nœuds ne possédant pas de fils) est associée une valeur de la variable cible (celle que l'on cherche à prédire : lettre ou non lettre).

4.2.2. Construction de l'arbre de décision

Plusieurs algorithmes sont possibles (ID3, C4.5, etc.). Dans le cas présent, nous avons implémenté l'algorithme ID3. Voici les différentes étapes de cet algorithme :

1. À la racine de l'arbre sont associés tous les textes extraits du fichier de données
2. La variable de segmentation (dans notre cas, un motif) qui maximise le gain d'information (minimisation de l'entropie de Shannon) pour ce nœud est sélectionnée
3. Une branche partant de ce nœud est créée pour chacune des valeurs possibles de cette variable de segmentation (deux branches : motif sélectionné présent ou non dans le texte)
4. Chacun des deux nœuds créés au bout de ces deux branches contient les textes correspondant à la valeur de la variable de segmentation de sa branche
5. On recommence à l'étape (2) pour ces deux nœuds

On arrête de diviser un nœud soit lorsque ce nœud contient uniquement des textes d'un même genre ; on attribue alors à ce nœud l'étiquette correspondante (lettre ou non lettre). Soit lorsqu'il n'y a plus de variables de segmentation disponibles. On attribue alors à ce nœud l'étiquette correspondant à la classe majoritaire (lettre ou non lettre) pour les textes de ce nœud.

L'entropie de Shannon (évoquée à l'étape (2) de l'algorithme ci-dessus) est représentée par la formule suivante : $H_S\left(L\middle|V\right) = -\sum_i P\left(X_i\right) \times \sum_k P\left(L_k\middle|X_i\right) \times \log P\left(L_k\middle|X_i\right)$ où

- $H_S\left(L\middle|V\right)$ est l'entropie de Shannon de ma variable à prédire L (lettre ou non) par rapport à la variable de segmentation V (un motif)
- $P\left(X_i\right)$ est la probabilité pour V d'avoir la valeur (motif présent ou non dans le texte)
- $P(L_k\middle|X_i)$ est la probabilité d'avoir (lettre ou non) sachant que V a la valeur

4.3. Phase de test et performance

Les différentes étapes pour attribuer un genre à un texte sont les suivantes :

1. On se place au niveau de la racine de l'arbre
2. Si le texte contient le motif correspondant au nœud courant, on suit la branche correspondante, sinon on suit l'autre branche
3. Si le nœud courant ne possède pas de fils, on attribue au texte le genre correspondant à ce nœud, sinon on retourne à l'étape (2)

La performance obtenue avec cette seconde méthode est également plutôt satisfaisante puisqu'elle permet de classer environ 93% des textes de l'ensemble de test de manière correcte. Il est à noter que, pour construire l'arbre de décision, tous les motifs (variables de segmentation) n'ont pas été utilisés. En effet, afin de réduire le temps de calcul nécessaire à la détermination de la variable de segmentation (motif) qui maximise le gain d'information pour un nœud donné, uniquement les quelques centaines de motifs les plus marqueurs des lettres et des non lettres (selon la liste de motifs construite avec la première méthode) ont été utilisés.

4.4. Analyse des données

L'application de cette seconde méthode fait également apparaître les motifs caractéristiques du genre épistolaire. On remarquera que les 1er motifs distingués sont identiques à ceux relevés dans la première méthode appliquée ; on retrouve, en effet, les formules wnn tAy.i Sa.t r spr.k iw.k … ou nfr snb.k.

Les arbres de décisions identifient une grande majorité des lettres. Il est important de noter que la méthode n'attribue jamais à tort le genre « lettre » à un document dont le genre n'est pas une lettre ; les quelques erreurs d'identification concernent toujours des lettres n'ayant pas été reconnues comme telles. Ainsi, 16 lettres n'ont pas été identifiées par la méthode des arbres de décision. Parmi celles-ci, quelques cas concernent une sous-catégorie du genre épistolaire : les lettres-littéraires et les lettres aux morts. Ces deux documents s'éloignent donc assez fort du genre épistolaire, la fonction des textes n'étant pas la même.

Parmi les lettres non-identifiées, on relève également quelques documents appartenant bien au genre épistolaire *stricto sensu*. Parmi ceux-ci, quelques cas de non reconnaissance peuvent aisément s'expliquer par l'état de conservation extrêmement lacunaire du document (lettres d'el-Hibeh) ; dans d'autres cas, quelques précisions sont nécessaires. Prenons deux exemples :

[1] La lettre conservée par l'O. Berlin P 10664 n'a pas été identifiée comme appartenant au genre épistolaire, pourtant, de nombreux indices laissent penser qu'il s'agit bel et bien d'une lettre :

- Ainsi, le document s'ouvre par un incipit du type NP n NP, incipit le plus couramment attesté dans les lettres du Nouvel Empire ; toutefois, le passage est partiellement lacunaire :

  Ex. : [sS ra-ms] n it-nTr imn-Htp [n tA Hw.t] nsw.t-bity wsr-mAa.t-ra-stp.n-ra anx wDA snb

« Le scribe ra-ms au père divin imn-Htp du temple du roi de Haute et Basse Égypte wsr-mAa.t-ra-stp.n-ra, VSF » (O. Berlin P 10664, r° 1-2)
Ep. Ramsès II – lettre

Si la fonction et le nom de l'expéditeur sont en lacunes, ces informations sont conservées pour le destinataire ; il s'agit d'Amenhotep dont la fonction est it-nTr « père divin ». Parmi les motifs retenus par l'arbre, l'un d'eux correspond partiellement à certains des lemmes rencontrés dans ce passage : n it-nTr sS Hw.t-nTr « au père divin et scribe du temple » ; ce titre se rencontre à plusieurs reprises dans le corpus épistolaire de la 21e dynastie dans la correspondance de deux personnages portant ce titre de fonction. Toutefois, ce motif n'a pas été identifié ici puisque, dans l'O. Berlin P 10664, le titre de fonction d'Amenhotep se limite à it-nTr, le syntagme sS Hw.t-nTr n'apparaît pas (n it-nTr sS Hw.t-nTr NP >< n it-nTr NP) ;

- La lettre se poursuit par une formulation caractéristique du registre des salutations : twi Hr Dd n mr.t-sgr « Je dis à Méretséger ». La divinité invoquée ici est donc Méretséger, déesse rarement mentionnée dans cette formulation. Dans cette expression caractéristique, les dieux Amon-Rê ou Amon-Rê-Horachti sont bien plus souvent mentionnés. Par ailleurs, le motif retenu dans l'arbre s'apparentant le plus à ce qu'on trouve dans notre lettre est le suivant : Dd n imn-ra « dire à Amon-Rê ». Toutefois, la divinité mentionnée est Amon-Rê ; aucun motif n'a été identifié avec la déesse Méretséger. Par ailleurs, on relève encore un motif retenu dans l'arbre correspondant aux premiers mots de cette formulation : r-nty twi Hr (marque d'ouverture du discours + je + préposition). Dans notre lettre, le marqueur d'ouverture du discours r-nty n'est pas présent, la formule s'ouvre directement par le pronom twi ; le motif n'a donc pas été reconnu comme discriminant.

  => d'où la nécessité d'intégrer des catégories sous spécifiées dans les motifs (partie du discours, catégories sémantiques, etc.) et la notion de variantes au sein d'un motif ;

- Le registre des salutations se poursuit par une tournure caractéristique, en lacune : twi Hr Dd n mr.t-sgr imy [ ], probablement à restituer twi Hr Dd n mr.t-sgr imy + verbe au subjonctif (anx, snb, ptr, wnn) suivi d'un pronom suffixe (.k) « Je dis à Méretséger de t'accorder [ ] » Ce type de formulation est tout à fait caractéristique du genre épistolaire ; elle fait, d'ailleurs, bel et bien partie des motifs apparaissant dans l'arbre de décision. Toutefois, le motif n'a pas été identifié dans ce cas, le passage, en lacune se limitant au verbe rdi (employé à l'impératif : imy). Dans la méthode des arbres de décision, il faudrait donc peut-être inclure dans le processus la notion de « lacune », le cas se présentant forcément régulièrement étant donné la nature du corpus.
- S'ouvre alors le propos même de la missive, introduit par un indicateur d'initialité, marque d'ouverture du discours, en lacune (r-Dd ?, Hna-Dd ?). La première proposition est partiellement en lacune ; seuls les premiers lemmes sont conservés : yA ix pAy[ ] « Que signifie ton/mon [ ] ». Cette construction (particule yA – interrogatif ix – adjectif possessif probablement suivi d'un verbe à l'infinitif) est attestée à plusieurs reprises dans le genre épistolaire. Ce motif n'a pas été identifié par la méthode des arbres.

[2] De même, la lettre conservée sur l'O. Černý 3+ O. Cambridge FM 1 n'a pas été identifiée comme appartenant au genre épistolaire, pourtant, de nombreux indices laissent penser qu'il s'agit bel et bien d'une lettre :

- Ce document s'ouvre par l'incipit NP Hr-nD-xr.t n NP, incipit régulièrement attesté depuis la 18e jusqu'à la 20e dynastie, plus souvent sur papyri :

  Ex. : sS twr Hr nD-xr.t n mw.t.f Smay.t n imn [ ]
  « Le scribe twr salue sa mère, la chanteuse d'Amon [NP] » (O. Černý 3 + O. Cambridge FM 1, r° 1)
  Ep. Ramsès II – lettre

  Cette tournure caractéristique des lettres figure parmi les motifs utilisés dans l'arbre. Toutefois, le motif retenu est nD-xr.t n sS « saluer le scribe » ; dans ce cas, le verbe nD-xr.t « saluer » est suivi de la préposition n « à » introduisant le titre de fonction sS « scribe », suite de lemmes attestée à plusieurs reprises dans le corpus épistolaire. Le motif identifié n'a donc pas été reconnu par la méthode puisque le destinataire de la missive est, dans ce cas, une **mère** portant le titre de « chanteuse d'Amon ».

  => À nouveau, cet exemple fait apparaître la **nécessité d'introduire la notion de « variante » dans l'identification d'un motif** (nD-xr.t n SN) ;

- Après l'incipit, on rencontre les formules de salutations habituelles :

  Ex. : m anx wDA snb m Hs.t imn-ra nswt nTr.w
  « En vie, prospérité, santé, dans les louanges d'Amon-Rê, roi des dieux » (O. Černý 3 + O. Cambridge FM 1, r° 2)
  Ep. Ramsès II – lettre

  Parmi les motifs identifiés comme discriminants, plusieurs correspondent à cette formulation. Toutefois, dans tous les cas, le motif retenu est toujours la formulation plus complète m Hs.t imn-ra nswt nTr.w pAy.i nb nfr « dans les louanges d'Amon-Rê, roi des dieux, mon bon maître ». Il n'a donc pas été identifié puisque notre document limite l'épithète d'Amon-Rê à nswt nTr.w « roi des dieux » ;

- Le registre de salutation se poursuit par une formule caractéristique vue plus haut : twi Hr Dd n + dieux « Je dis à + Dieux ». Dans ce cas, l'expression est en lacune ; le motif identifié dans l'arbre n'a donc pas été reconnu ;

- Dans cette même proposition apparaît une épithète fréquemment associée à une divinité solaire : Dieu + iw.f Hr wbn [ ], probablement à restituer NP iw.f (Hr) wbn Htp ou NP iw.f (Hr) wbn iw.f (Hr) Htp « Dieu, quand il se lève et se couche ». Aucun motif correspondant n'a été retenu dans l'arbre afin de scinder deux groupes de textes ;

- Le registre de salutations se poursuit encore par trois formules caractéristiques du genre épistolaire :

  Ex. : tw[i Hr Dd n] imn mw.t xnsw (…) imy snb.T imy wnn.T m Hs.t tA dhn.t imn.t tA Hnw.t imy p[tr.i ss]nb.tw
  « Je [dis à] Amon, Mout et Khonsou (…) qu'ils fassent en sorte que tu sois en bonne santé, qu'ils te placent dans la faveur de la falaise occidentale, la maîtresse, qu'ils

fassent en sorte que [je te voie] en bonne santé » (O. Černý 3 + O. Cambridge FM 1, r° 2-4)
Ep. Ramsès II – lettre

Ces diverses formulations figurent parmi les motifs répertoriés dans l'arbre de décision : ssnb « guérir » et rdi snb.k « faire en sorte que tu sois en bonne santé », toutefois, aucun d'eux n'est reconnu dans ce document. Dans le premier cas, il s'agit d'une erreur de la part de l'encodeur du texte qui a lu snb.tw et non ssnb.tw. Dans le second cas, le motif retenu est rdi snb.k « faire que tu sois en bonne santé » où le verbe rdi introduit le verbe snb au subjonctif suivi du pronom suffixe de la 2e pers. masc. sg. Or, dans notre lettre, on trouve la formule rdi snb.T ; le pronom est, cette fois, le pronom suffixe de la 2e pers. fém. sg, le destinataire de la lettre étant une femme.

Ce cas témoigne à nouveau de la nécessité d'intégrer la notion de motifs multidimensionnels ; il s'agit, en effet, de la même formule ;

- Notre document conserve également quelques expressions caractéristiques du genre épistolaire. Ainsi, la formule xy qd.k « Comment vas-tu » est typique du genre épistolaire ; elle n'est jamais employée dans les autres genres. Il faut toutefois préciser qu'elle apparaît peu et est confinée aux lettres de la 18e dynastie et du début de la 19e dynastie. Ce motif n'a pas été identifié comme discriminant par la méthode de l'arbre. De même, un peu plus loin dans la lettre apparaît l'expression ix-di.T Hr.T « Puisses-tu être attentive ». À nouveau, ce motif n'a pas été identifié comme critère de discrimination entre deux groupes de textes.

5. Conclusion

Les méthodes présentées apportent toutes deux des résultats similaires plutôt satisfaisants au niveau des performances de classification (plus de 90% de classifications correctes grâce à ce seul critère). Cependant, la seconde méthode présente l'avantage majeur de rendre explicite les éléments (présence ou non de motifs dans le texte à classer) qui amènent à la classification d'un texte comme étant ou non une lettre.

Au terme de ce travail, il nous est apparu que plusieurs points pouvaient encore être développés. Ainsi, il s'avérerait utile d'intégrer des catégories sous-spécifiées dans les motifs (partie du discours, catégories sémantiques, etc.) afin de détecter les variantes apparaissant au sein d'un même motif. De même, étant donné la nature parfois très fragmentaire du corpus néo-égyptien, il nous semble indispensable de tenir compte de la présence de lacunes au sein des motifs. Dans notre cas, un motif comme « A B C D E » sera distingué du motif « A B lacune D E » (qui sera en fait considéré comme étant « A B D E ») ; or il s'agit bien du même motif.

Dans la continuité de ce travail, nous envisageons d'appliquer ces méthodes à d'autres genres textuels (documents royaux, textes juridiques, etc.) du corpus néo-égyptien. De même, notre travail se limite à l'identification des textes d'un seul genre par rapport au reste du corpus ; il pourrait donc s'avérer intéressant de voir ce que donnerait une application des méthodes exposées dans cette contribution sur plusieurs genres simultanément. On aurait alors une classification qui ne serait plus limitée à une dichotomie d'un genre par rapport aux autres, mais bien de chacun des genres par rapport aux autres. Ce type d'approche pourrait bien augmenter

les performances de classification pour chacun des genres. En effet, l'opposition d'un genre à un ensemble de genres bien déterminés pourrait alors s'avérer plus efficace que l'opposition d'un genre au reste du corpus pris comme un tout.

Références

Bakir A. (1970). *Egyptian Epistolography from the Eighteenth to the Twenty-First Dynasty*, Le Caire (= *BdE* 48).

Longrée D., Luong X., Mellet S. (2008). *Les motifs : un outil pour la catégorisation topologique des textes*, dans Heiden S., Pincemin B. (éd.), *Actes des JADT 2008, 9èmes Journées internationales d'Analyse statistique des Données Textuelles*, Lyon, 12-14 mars 2008, pp. 733-744.

Mellet S., Luong X., Longrée D., Barthélémy J.-P. (2009). *Représentations du texte pour la classification arborée et l'analyse automatique de corpus. Application à un corpus d'historiens latins*, *Mathematics and Social Sciences*, 187, pp. 107-121.

Polis St., Winand J., Honnay A.-C. (2010). *The Ramses Project. Review and Perspectives*, Liège (= *Aeg. Leod.*).

Rosmorduc S., Polis St. &Winand J. (2009).*Ramses. A New Research Tool in Philology and Linguistics*, dans N. Strudwick (ed.), *Information Technology and Egyptology*, Piscataway (N.J.), *Proceedings of the XXIst Table Ronde « Égyptologie et Informatique »*, pp. 155-164.

Winand J., Polis St., Rosmorduc S. (sous-presse).*Ramses. An Annotated Corpus of Late Egyptian*, dans P. Kousoulis (éd.), *Proceedings of the Xth International Association of Egyptologists Congress* (Rhodes, Mai 2008), Louvain.