Textométrie hiéroglyphique

[Conte du naufragé]

André Salem, Romuald Schummer
salem@msh-paris.fr, schummer2001@yahoo.fr

They did not know it was impossible, so they did it !
Mark Twain[1]

Résumé : A partir d'un texte hiéroglyphique et de ses translittérations sur un support informatisé, les méthodes textométriques permettent d'explorer directement des récurrences textuelles contenues dans le corpus. Le repérage de séquences répétées dans le texte original ouvre une voie textométrique à l'étude des procédés narratifs à l'œuvre dans le récit. La constitution d'un bitexte constitué du texte original et de sa traduction française alignée au niveau du verset permet d'étudier l'activité de traduction réalisée à partir des textes originaux.

Mots clés : textométrie, hiéroglyphes

L'activité d'*exploration* recèle bien des dangers pour ceux qui s'aventurent sans préparation dans des contrées qu'ils n'ont pas pris le temps de connaître, au moins par les récits de gens qui en sont revenus sains et saufs. En abordant l'exploration textométrique de textes fixés sur parchemin il y a plusieurs millénaires, après avoir connu une existence que l'on peut supposer aussi longue sous forme de poèmes transmis oralement de générations en générations, nous avons pleinement conscience de ne pas avoir préparé notre voyage avec autant de soin qu'il aurait été utile de le faire.

D'un autre côté, nous disposons aujourd'hui d'un corps de méthodes et d'outils textométriques éprouvés sur de très nombreux textes, écrits dans des langues extrêmement diverses. Ces méthodes ont montré qu'en s'appuyant sur la forme matérielle du texte et en y projetant un éclairage quantitatif, il était possible d'y repérer de *faits textuels* de répartition ou de répétition que les spécialistes formés aux sciences humaines, plus naturellement enclins lors de leurs lectures cursives à en extraire ce qui fait sens pour eux, en s'appuyant sur l'érudition acquise à leur contact, risquaient de négliger.

L'intuition textométrique souffle que cet éclairage devrait également prouver son efficacité sur les séquences de caractères hiéroglyphiques[2] que les systèmes informatiques modernes permettent désormais de gérer.

1 Le contexte de la recherche

[1] Citation placée en exergue sur le site du *Projet Rosette* (http://projetrosette.info/) sur lequel nous avons recueilli l'essentiel des ressources informatisées qui nous ont permis de réaliser cette étude.

[2] Du grec ἱερογλύφος / *hieroglúphos*, composé de ἱερός / *hierós* sacré et γλύφειν / *glúphein* graver.

Dans ce qui suit, notre projet sera double. Nous aimerions, en premier lieu, attirer l'attention des différents spécialistes de l'étude des textes hiéroglyphiques sur l'efficacité des méthodes textométriques et sur les possibilités d'investigation nouvelles qu'elles ouvrent aux chercheurs dans le domaine des études égyptologiques. Par ailleurs, il nous semble que cette première application de méthodes textométriques, souvent éprouvées sur des corpus de textes rédigés dans des langues modernes, à des textes qui relèvent d'un système d'écriture très différent peut permettre du même coup à la communauté des études textométriques de prendre un recul utile par rapport au corps de méthodes qu'elle met régulièrement en œuvre sur les corpus de texte qui retienne son atention.

2 Le système d'écriture hiéroglyphique[3]

Les textes hiéroglyphiques sont en fait composés de phrases regroupant des mots écrits à l'aide de signes-images. Il n'y a pas de ponctuation et, comme c'est le cas pour la plupart des systèmes d'écriture de l'Antiquité, les mots ne sont pas séparés par des espaces. L'ordre dans lequel le texte doit être lu varie d'une inscription à l'autre (gauche-droite, droite-gauche, haut-bas, parcours boustrophédon, etc.).

Le système d'écriture hiéroglyphique permet et encourage même, à des fins esthétiques, des modifications de la séquence linéaire du texte. Les signes sont dessinés à l'intérieur d'un carré imaginaire qu'on appelle *cadrat*. Il sont parfois regroupés en un empilement méthodique, certains signes pouvant être associés ou superposés par rapport à d'autres.

2.1 Classification des hiéroglyphes par leur fonction

On peut classer les signes en trois classes principales :

- *idéogrammes :* certains signes sont utilisés pour coder le nom de l'être, de l'objet ou de l'action qu'ils représentent. L'image d'un taureau **Ä** permet la référence à cet animal, celle d'un plan de maison **O** est utilisée pour signifier *maison*. L'image d'une voile gonflée par le vent ¶ est utilisée pour faire référence au *vent*.
- *phonogrammes :* d'autres signes sont principalement utilisés pour représenter un son. L'image d'un serpent **œ**, correspond plus ou moins au groupe phonique « dj », celle d'une bouche **'** que l'on prononce « er » sert à représenter la lettre « r », etc.
- *déterminatifs :* pour réduire le nombre des ambiguïtés dues à l'homonymie, on utilise des déterminatifs placés en fin de mot qui ne se prononcent pas. Ainsi, dans cette fonction, l'homme assis **!** détermine la séquence qui précède comme : occupations masculines, noms propres, etc.

Notons qu'un même signe peut avoir des fonctions différentes en fonction du contexte dans lequel il est utilisé.

2.2 Translittérations modernes

En 1927, un siècle après la classification de Champollion, Gardiner propose une classification portant sur les quelques 740 hiéroglyphes, les plus courants. Chacune des 26 catégories de cette classification est symbolisée par une lettre. A l'intérieur de chaque catégorie les hiéroglyphes sont numérotés à partir de 1. Le code A1 correspond, par exemple, au signe **!** (homme assis), le code A2, au signe **#** (homme assis portant la main à la bouche), etc.

Pour les translittérations modernes, on utilise de plus en plus les prescriptions du *Manuel de codage* (dorénavant MdC) adoptées en 1988 par une grande partie de la communauté des

---

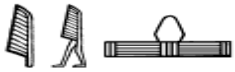
[3] Pour cette présentation des grandes lignes du système d'écriture hiéroglyphique, nous avons utilisé l'ouvrage publié par le ministère français de la culture à l'occasion de l'exposition *Naissance de l'écriture, cunéiforme et hiéroglyphes* - Galeries nationales du Grand Palais, Éditions de la réunion des musées nationaux, Paris, 1982.

égyptologues, qui permettent de transcrire les textes hiéroglyphiques en utilisant à la fois les codes de Gardiner et les translittérations de certains phonogrammes les plus courants.[4]

2.3 Codage informatique des écrits hiéroglyphiques

Le codage informatique moderne s'appuie notamment sur ces dernières méthodes de translittération pour stocker les textes initialement composés sous forme hiéroglyphique. A cette translittération vient souvent d'ajouter un découpage en *mots*. Chaque séquence reconnue comme un mot est précédée par un blanc et/ou caractère informatique particulier, les différents morphèmes grammaticaux étant systématiquement isolés par d'autres caractères[5].

Ainsi, la séquence de signes :



dont le codage dans la liste Gardiner est : **M17 M18 R4** sera notée, dans ce système de codage, à partir de ses valeurs phonétiques : **i ii Htp**.

Dans les transcriptions que nous avons utilisées, les codes « : » et « * » permettent respectivement de transcrire la superposition et la juxtaposition de deux signes. Le groupe de signes :

sera codé : **p*t:pt**, d'après ses valeurs phonétiques ou : **Q3*X1:N1** d'après les codes de la liste de Gardiner (association des signes Q3 et X1 dessinée au-dessus du signe N1).

2.4 Transcriptions, translittérations, traductions

Partant d'un texte hiéroglyphique, ont peut *générer*, en utilisant dans chaque cas des règles dont le degré de formalisation varie selon l'objectif fixé, d'autres *textes* qui permettront à des individus moins versés dans la lecture hiéroglyphique de mieux saisir tel ou tel aspect de la signification ou de la prononciation du texte :

- *une translittération :* substitue à chaque **graphème** d'un système d'écriture un **graphème** ou un groupe de graphèmes d'un autre système, indépendamment de la prononciation. Si les règles de translittération sont explicites et réversibles, il est possible de reconstituer le texte original à partir du résultat de la translittération.
- *une transcription :* substitue à chaque **phonème** d'une langue un **graphème** ou un groupe de graphèmes d'un système d'écriture.
- *une traduction :* tente de restituer dans une autre langue le **sens** contenu dans le texte original. Dans la pratique, les traducteurs choisissent entre plusieurs options dont certaines visent à rester au plus près du texte original pour le *trahir* le moins possible, alors que d'autres prennent, au contraire, le parti de placer la traduction dans un cadre socio culturel familier au lecteur, afin de faciliter au maximum sa perception du texte original.

Comme on le comprend, les *translittérations* et les *transcriptions* peuvent posséder, sous certaines conditions, la propriété de **réversibilité**. Tel est le cas, par exemple, si chaque état

---

[4] Cf. *Manuel de codage des données pour textes hiéroglyphiques sur ordinateur*, *consultable par exemple sur le site : http://projetrosette.info/page.php?Id=205.*

[5] On trouvera, au tableau 2, l'exemple d'un texte hiéroglyphique muni de sa codification dans un codage de ce type.

du texte est accompagné des règles de translittération qui, associées à ce texte, permettent de reconstituer l'état original. Dans ce cas, on peut grosso modo considérer, au plan textométrique, chacune des translittérations obtenues comme des ressources équivalentes au texte original. Comme on le conçoit aisément, cette propriété est rarement associée aux traductions effectuées d'une langue à une autre. Les traductions ne suffisent pas, dans le cas général, à reconstituer de manière univoque le texte original.

2.5 Segmentation en mots

Comme nous l'avons signalé plus haut, la tradition d'écriture hiéroglyphique ne sépare pas systématiquement par des blancs les différents mots qu'un lecteur égyptologue peut identifier dans le texte. Pour venir à bout de cette tâche, il est possible de s'appuyer sur le repérage de certains signes (ex : les déterminatifs) qui apparaissent prioritairement en fin de mot. Cependant, les spécialistes s'accordent sur le fait qu'une solide connaissance de la langue est nécessaire pour découper un texte hiéroglyphique en mots[6].

2.6 Ressources hiéroglyphiques en ligne

Un certain nombre de translittérations, et tout particulièrement celles qui permettent de redessiner les signes hiéroglyphiques originaux à partir des translittérations de type Gardiner, peuvent être confiées à des procédures informatiques. L'utilisation de telles procédures permet du même coup de vérifier le bon encodage du texte translitteré et de garantir l'homogénéité de la translittération elle-même.

Plusieurs sites web proposent des procédures capables d'effectuer automatiquement cette opération[7]. A partir du texte translitteré, ces procédures restituent des images qui permettent de vérifier visuellement la conformité de la translittération réinterprétée au texte d'origine. Les procédures réunies sur le site du ***Projet Rosette*** permettent, de plus, de faire le lien, pour chaque signe hiéroglyphique, avec toute une série de renseignements de type dictionnairique qui concernent : ses variantes scripturales, sa prononciation, sa signification globale, ses différentes significations en contexte, etc.

Ces possibilités de transcriptions automatiques fiables permettent de considérer les corpus de textes hiéroglyphiques translittérés comme des bases de données textométriques susceptibles de servir de point de départ à des traitements textométriques dont les résultats pourront également être translittérés sous leur forme hiéroglyphique originale.

3 Le corpus *Naufragé*

Le ***Conte du Naufragé*** est l'un des textes importants de la littérature de l'Égypte ancienne parvenus jusqu'à nous. Des versions électroniques du texte hiéroglyphique original, composé de 190 versets, ainsi que des traductions, des transcriptions et des translittérations destinées à permettre la conservation de ce texte sur des supports informatisés peuvent être aisément localisés sur différents sites consacrés à l'égyptologie.[8] Le *Conte du naufragé* à donné lieu à

---

[6] De cette certitude partagée par les égyptologues, on peut inférer sans risque de se tromper qu'à l'instar de ce qui se passe pour les textes écrits en d'autres langues, tout découpage d'un texte hiéroglyphique en mots et a fortiori toute tentative de rattacher systématiquement chacun des mots découpés dans la chaîne textuelle à des unités dictionnairiques plus génériques (lemmatisation) sera susceptible de prêter le flan à des critiques qui feront valoir des interprétations du texte ou des arguments de grammairiens conduisant à des découpages et ou à des regroupements différents.

[7] Pour cette étude, nous avons eu recours à l'ensemble des procédures réunies sur le site du Projet Rosette : http://www.projetrosette.info.

[8] La version électronique du texte hiéroglyphique du *Conte du naufragé* que nous avons utilisée pour cette étude a été téléchargée à partir du site du Projet Rosette.

de nombreuses études de caractère littéraire portant essentiellement sur la structure extrêmement remarquable du récit[9].

==== Le conte du naufragé ====

Le papyrus : La seule version de ce conte qui nous soit parvenue est consignée sur un papyrus hiératique[10]. Le document a été découvert dans les réserves du Musée de l'Ermitage, à Saint-Pétersbourg à la fin du 19ème siècle de notre ère. Les historiens qui ont pu faire des rapprochements avec d'autres textes fixés sur papyrus à la même époque pensent que le document a été établi il y a environ 4 000 ans.

Il n'est pas possible d'estimer avec précision la date de la création du récit lui-même. Bien avant sa fixation sous forme écrite, ce texte z pu circuler sous forme d'un récit poétique transmis oralement, sans altération majeure, de générations en générations pendant une très longue période. Le texte peut avoir été traduit ou fortement inspiré par un texte préexistant transmis oralement ou fixé sur un document rédigé dans une autre langue.

L'histoire : Pour rassurer un jeune supérieur, inquiet d'avoir à rencontrer prochainement son suzerain, un vieux serviteur lui raconte qu'embarqué sur un navire il a été victime d'un naufrage qui l'a fait échouer sur une île habitée par un serpent géant. Sa frayeur dissipée, il a raconté son histoire au serpent. Puis le naufragé a écouté l'histoire du serpent, lui même victime de malheurs qui ont abouti à la destruction de sa propre famille, lors d'une période précédente. A l'issue de cette rencontre, le serpent a couvert le naufragé de présents et lui a prédit qu'il vivrait heureux parmi les siens. Le jeune supérieur écoute avec attention ce récit qui ne dissipe cependant pas ses propres craintes.

La critique : Plusieurs critiques modernes ont souligné la composition originale de ce récit. Plusieurs conteurs y enchâssent à tour de rôle des récits personnels ainsi que des commentaires sur les faits qu'ils relatent. On note des symétries dans la manière dont sont agencées les différentes parties du conte. A la description du voyage d'aller correspond celle d'un retour, aux frayeurs initiales, des surprises agréables, etc.

---

[9] Cf., par exemple, D. Benoît, Le conte du naufragé dans le cycle : Les grands textes de l'Égypte ancienne. http://www.thotscribe.net/docs/2004_2005/conte_naufrage.pdf.

[10] L'écriture *hiératique* constitue une forme simplifiée de l'écriture hiéroglyphique permettant d'écrire plus rapidement.

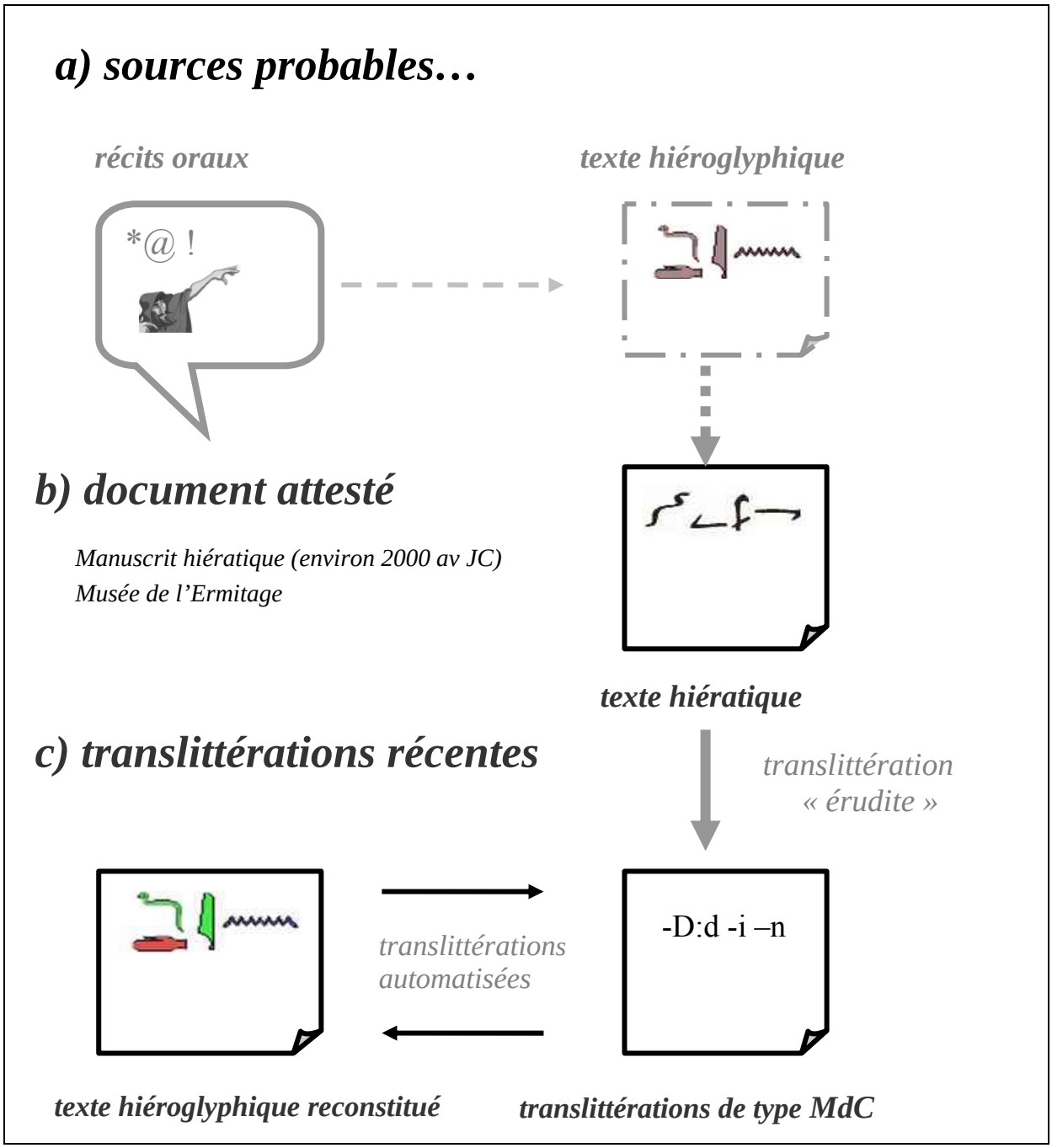


Figure : 1

Le conte du naufragé :

sources probables, documents attestés et translittérations modernes

On a rassemblé, sur la figure 1, différents états du récit qui ne nous est parvenu que sous forme d'un papyrus hiératique (section b). Les états antérieurs de ce récit, dont l'existence est probable, sont mentionnés en gris (section a). La dernière section (section c) regroupe les versions informatisées du texte sur lesquelles nous avons pu travailler effectivement.

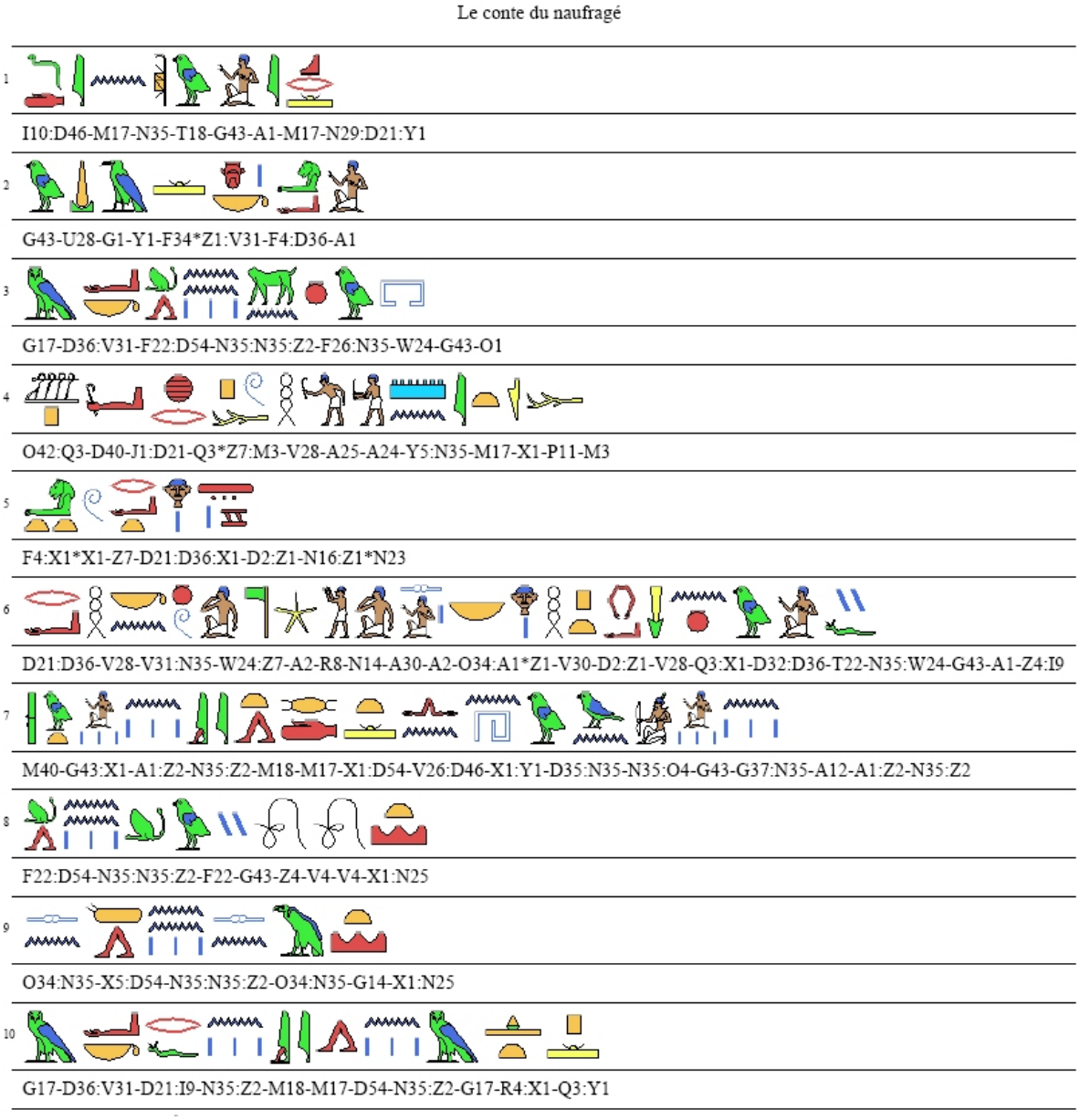


Figure 2

Le début du *Conte du naufragé*
suivi de sa transcription en codes Gardiner

Dans l'état actuel des logiciels de traitement textométrique, il n'est pas envisageable d'appliquer directement les procédures textométriques destinées à permettre des mesures sur le texte aux séquences de signes hiéroglyphiques. Cependant, comme nous l'avons signalé plus haut, les translittérations obtenues à partir des codages de type Gardiner constituent des équivalents extrêmement fidèles de leur source hiéroglyphique.

Dans ce qui suit, nous tenterons de montrer comment des mesures textométriques portant en fait sur les séquences originales du texte hiéroglyphique peuvent être réalisées à partir des translittérations de type Gardiner archivées sur support informatique.

4 Approches textométriques du corpus *Naufragé*

Pour soumettre un texte à des traitements textométriques, il est nécessaire de déterminer deux systèmes complémentaires : un système de *contenants*, parties du texte qui vont être soumises à des comparaisons textométriques et un système de *contenus,* unités textuelles (habituellement : *mots, graphèmes*, etc.) dont on s'attachera ensuite à recenser les occurrences au sein de chacune des parties du texte.

A partir du décompte des occurrences des unités-contenus à l'intérieur des contenants, les méthodes textométriques produisent des jugements quantitatifs qui peuvent ensuite être interprétés en terme de variations dans l'usage du vocabulaire.

Nous avons jugé utile, dans ce qui suit, de faire figurer, en regard des calculs effectués à partir du texte hiéroglyphique, des calculs similaires réalisés à partir de la traduction française du *Conte du naufragé*. On peut voir, sur le tableau 2, un extrait de chacune des deux versions du texte qui constituent ensemble ce que l'on appelle un *corpus aligné multilingue*. L'alignement a été réalisé, ici, au niveau du verset. A côté des calculs que l'on peut effectuer à partir de chacun des volets pris isolément, les investigations multilingues permettent d'effectuer des rapprochements entre fragments du corpus aligné et de mieux analyser l'activité de traduction effectuée entre les deux versions du texte.

4.1 Découpages du corpus

Le document original se présente sous forme d'un texte découpé en 190 lignes, que nous appellerons ici des ***versets***. Une même phrase du texte (un même mot ?) peut se retrouver transcrite à cheval sur la fin d'un verset et sur le début du verset suivant. Nous avons numéroté les versets de 1 à 190 en faisant précéder le verset x de la balise <v=x>.

Tableau : 1
Partition en douze fragments du corpus ***Naufragé***

Partie	occurrences	formes	hapax	fmax	forme
01Intro	313	78	35	31	n
02VoyageEtNaufrage	277	81	37	21	n
03IleDuKa	251	73	40	28	n
04LeSerpent	434	90	39	39	n
05RecitNaufAuSerp	374	91	40	29	n
06DiscduSerpent1	224	61	28	23	n
07RecitduSerpent	270	73	37	31	n
08NaufetSerpent	354	91	41	30	n
09DiscDuSerpent2	597	115	60	60	n
10Retour	214	59	21	31	n
11Epilogue	153	56	29	13	n

Dans le document original, certains groupes de versets sont précédés d'une courte séquence de signes mise en valeur par une coloration rouge qui semble marquer le début d'une nouvelle

partie du récit et suggérer un découpage du texte en parties[11]. Ce découpage provisoire, dont il faut noter que nous ignorons l'origine exacte, ne constituera pas pour nous une donnée indépassable. Nous nous appuierons cependant sur ce découpage pour effectuer une première comparaison à partir des différents fragments du texte.

Nous avons transcrit cette division qui aboutit à une partition du texte en douze fragments par des balises de type <D=y> où *y* varie de 1 à 12. Le tableau 3 fournit les principales caractéristiques lexicométriques calculées pour chacun des fragments.

On trouve au tableau 3 un état qui présente le début de chacun des deux volets du corpus munis des balises qui permettent de distinguer les versets et les regroupements thématiques.

4.2 Les unités de décompte

La question de la détermination des unités les plus aptes à servir de base aux décomptes textométriques a longtemps agité les communautés de chercheurs confrontées aux corpus textométriques[12]. Nous avons signalé que, dans le cas des corpus hiéroglyphiques, la détermination des frontières de mots constituait une tâche hors de portée pour les traitements automatisés. Nous consacrerons l'essentiel de cette première étude au repérage automatique des répétitions contenues dans le texte. Pour effectuer cette tâche, nous allons commencer par considérer le système des unités de décompte constitué par les différents signes hiéroglyphiques.

```
<D=01Intro>
<v=001> -D:d -i -n -Sms -w -A1 -i -q:r:Y1 -w -DA -A -Y1 §
<v=002> -ib*Z1:V31A -HAt:a -A1 -m -a:V31A -pH:D54 -n:n:Z2 §
<v=003> -Xn:n -nw -w -pr -Ssp:p -a -x:r -p*W:xt §
<v=004> -H -A25 -A24 -mn:n -i -t -P11 -xt -HAt:t*t -W -r:a:t §
<v=005> -Hr:Z1 -tA:Z1*N23 -r:a -H -V31A:n -nw:W -A2 -nTr -dwA §
<v=006> -A30 -A2 -z:A1*Z1 -nb -Hr:Z1 -H -p:t -D32:a -sn -n:nw -w -A1 -y:f §
<v=007> -iz -w:t -A1 -Z2 -t:n:Z2 -ii -i -t:D54 -aD:d -t:Y1 -D35:n §
<v=008> -n:h -w -wr:n -mSa -A1:Z2 -n:Z2 -pH:D54 -n:n:Z2 §
<v=009> -pH -w -y -wA -wA -t:xAst -z:n -X5:D54 -n:n:Z2 §
<v=010> -z:n -mwt -t:xAst -m -a:V31A -r:f -n:Z2 -ii -i -D54 -n:Z2 §
<v=011> -m -Htp:t -p:Y1 -tA:N23*Z1 -n:Z2 -pH:D54 -n:Z2 -sw -W §
```

```
<D=01Intro>
<v=001> un excellent suivant dit alors : apaise §
<v=002> ton coeur, prince ! vois, nous avons atteint §
<v=003> la résidence. le maillet est saisi et §
<v=004> le poteau d'amarrage est frappé, l'amarre de proue ayant été portée §
<v=005> à terre ; les prières sont dites, le dieu a été remercié §
<v=006> et chaque homme embrasse son semblable, §
<v=007> car notre équipage est revenu sain et sauf, sans §
<v=008> perte pour notre troupe. nous avons atteint §
<v=009> les confins de ouaouat, après avoir doublé §
<v=010> senmout. vois donc, nous revenons  §
<v=011> en paix, notre pays, nous l'avons atteint. §
```

Tableau 2

Le corpus multilingue aligné ***Naufragé***
a) le début du poème codé selon les normes MdC
b) la traduction française de cet extrait

[11] Pour effectuer ce découpage, nous nous sommes efforcés de suivre les indications du manuscrit original qui ont donné lieu à l'insertion d'intertitres (rédigés par les éditeurs français du manuscrit) sur le site sur lequel nous avons récupéré le texte orignal.

[12] Sur ces questions on consultera, par exemple, [Muller 1963] et [Brunet 2000].

Pour mettre en œuvre ce choix, il nous suffira de considérer, dans le cadre de cette première expérience, les signes d'association (*) et de superposition (:) comme des caractères isolant les différents signes réunis dans un même cadrat. Cette option s'appuie sur l'affirmation trouvée dans les travaux que nous avons pu consulter, que l'habitude de superposer et d'associer différents signes hiéroglyphiques dans un même cadrat prend souvent sa source dans des considérations d'ordre esthétique. Si cette hypothèse est vraie, on peut s'attendre à ce que les séquences de signes ayant donné lieu au regroupement graphique en un même cadrat composite soient traitées de la même manière aux différents endroits du texte dans lesquels elles apparaissent. Notons que la prise en compte du texte sur support informatisé nous permet de vérifier systématiquement cette hypothèse par l'utilisation de la méthode textométrique de base que constitue l'établissement de *concordances*.

4.3 Principales caractéristiques textométriques

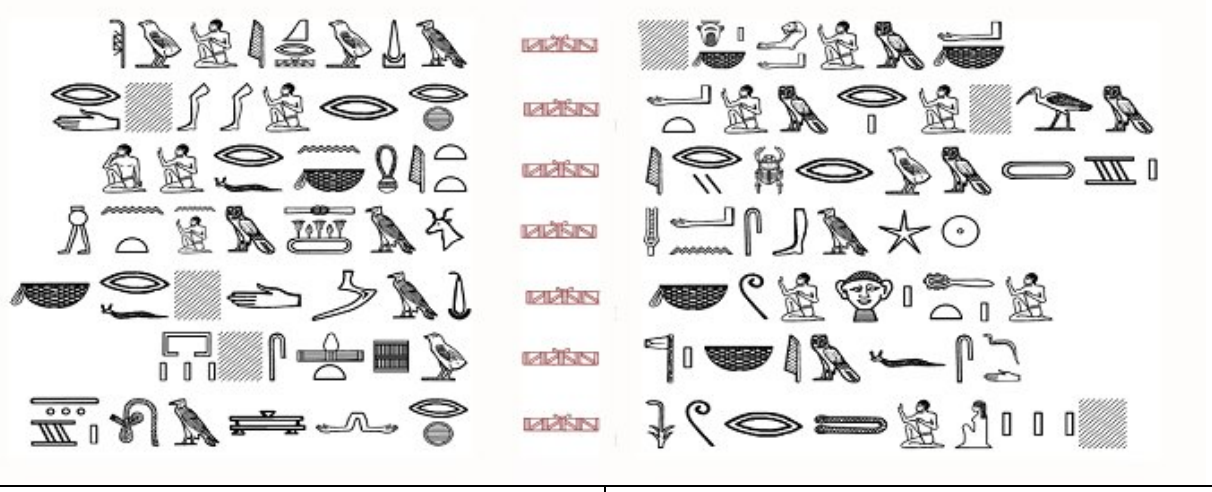
Le dépouillement des deux volets du corpus parallèle amène les caractéristiques lexicométriques que l'on trouve au tableau 3. Ces caractéristiques ne sont pas directement comparables car elles signalent avant tout des différences notables dans les systèmes d'écriture, compte tenu des normes de dépouillement que nous avons utilisées. Dans le cas du volet français du texte, la segmentation s'est faite sur des unités lexicales qui correspondent plus ou moins aux mots de la langue. Dans le cas du corpus hiéroglyphique, la segmentation a abouti à isoler des unités plus ténues qui entrent dans la composition des mots (lettres, phonèmes, morphèmes, déterminants). Les caractéristiques lexicométriques calculées sur chacun des volets du corpus portent la trace de cette importante différence. Les différents modes de segmentation retenus expliquent à eux seuls : d'une part le plus grand nombre d'occurrences et la fréquence maximale nettement plus élevée dans le volet hiéroglyphique, de l'autre, le plus grand nombre de formes et d'hapax dans la traduction française du texte.

Tableau : 3
Principales caractéristiques textométriques
pour les deux volets du corpus ***Naufragé***

	Hiéroglyphes	Français
Nombre d'occurrences	3 741	1 745
Nombre de formes	248	541
Nombre d'hapax	89	316
Fréquence maximale	336	77
forme	n	de

4.5 Concordance d'un signe

Lorsqu'on désire étudier la signification d'une unité textuelle dans l'ensemble d'un corpus ou examiner chacun de ses contextes particuliers d'utilisation, la possibilité de rassembler sur un même document toutes les occurrences d'une forme donnée, accompagnée d'un contexte minimal, constitue l'un des avantages les plus appréciables offerts par la prise en compte d'un corpus informatisé.



Signe	***Signification***
 Code Gardiner : **Y1** EGPZ : 58328 (e3d8) GlyphBasic : 4-242 Transliteration : mDA.t / dmD / dmd	écriture, abstraction ***Description*** rouleau de papyrus scellé (var.Y2) ***Commentaire*** : - idéogramme dans mDAt 'rouleau de papyrus' - déterminatif dans les termes liés à l'écriture ou aux notions abstraites

Figure : 3
Extrait d'une concordance réalisée à partir de la forme Y1 *écriture*
(les carrés gris signalent un changement de verset)

Comme on l'a souligné plus haut, dans le cas d'une translittération chacune des occurrences d'une même unité textuelle reçoit un codage identique. Dans notre cas, chacun des signes hiéroglyphiques reçoit un code identique. Pour réaliser la concordance du signe  que l'on peut voir sur la figure 3, nous avons commencé par réaliser une concordance portant sur les occurrences de la forme Y1 dans le fichier translitteré. Les lignes de contexte générées par le module de concordance ont ensuite été soumises à l'éditeur Rosette[13] qui a rétabli leur forme hiéroglyphique originale.

Les états ainsi obtenus permettent d'examiner sous forme *visuelle* l'ensemble des emplois d'une même unité de segmentation dans un corpus de textes hiéroglyphiques.

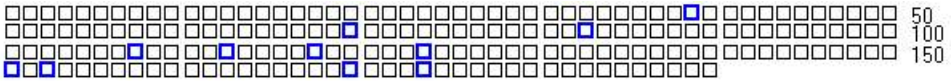
4.4 Explorations multilingues

Le fait de disposer d'une traduction alignée du texte que l'on étudie se révèle d'une grande utilité pour explorer un texte rédigé dans une langue que l'on ne domine pas. Les méthodes textométriques permettent d'établir des liens entre certaines des unités textuelles qui sont en rapport de traduction au sein d'un bitexte aligné.

Ainsi, par exemple, on peut constater que le terme *île* apparaît onze fois dans le volet français du corpus. Pour tenter de trouver des termes qui correspondent à ce terme dans le volet hiéroglyphique du corpus, on commence par sélectionner les versets qui contiennent la forme île dans le volet français (figure 4a).

[13] Le site Projet Rosette offre un éditeur *en ligne* qui traduit sous forme hiéroglyphique les séquences de signes translittérés qui lui sont fournies par le biais d'un interface web.

4a : versets du volets français contenant la forme *île*



4b : sélection des versets du volet hiéroglyphique correspondant à des versets du volet français contenant la forme *île*

Figure : 4
Extraction de termes en rapport de traduction à partir d'un bitexte

On commence par repérer les sections du volet français dans lesquelles apparaît le terme *île*. Pour chacune de ces sections, on peut localiser, dans le volet hiéroglyphique, une section correspondante laquelle est susceptible de contenir un terme en rapport de traduction avec cette forme lexicale. Le calcul des spécificités (formes surreprésentées) dans la zone du volet hiéroglyphique ainsi mise en évidence nous indique que la séquence de signes -iw:N23*Z1 apparaît 11 fois dans le corpus. L'éditeur du site Rosette nous fournit la forme hiéroglyphique originale de cette translittération et nous informe que ce signe complexe se traduit bien en français par le nom commun *île*[14].

	N18:N23*Z1	iw	**nc** : île

4.6 L'accroissement du vocabulaire

La figure 5 montre la courbe du vocabulaire réalisée pour le volet hiéroglyphique du corpus ***Naufragé***. La partition du corpus en fragments a été matérialisée sur ce graphique par des lignes verticales qui marquent chacune le début d'un des douze fragments du corpus.

Certains fragments sont caractérisés par des portions presque horizontales de la courbe d'accroissement. Cette circonstance peut s'expliquer par le fait que ces fragments sont le siège de répétitions de signes hiéroglyphiques déjà utilisés dans des fragments précédents.

La seconde courbe rend compte de l'apparition des *hapax* (formes qui ne trouvent qu'une seule occurrence dans le corpus).[15] Dans les dépouillements textométriques pratiqués à partir du découpage du texte en *mots*, on a pu remarquer que, loin de constituer une exception, la

---

[14] Ce type de procédure a été analysé par Maria Zimina dans sa thèse, cf. [Zimina 2004]. Les versions actuelles de Lexico3 (à partir de la version 3.5.9) permettent d'interroger chacun des volets d'un corpus parallèle à partir d'un sélection effectuée sur l'autre volet.

[15] Dans la longue tradition des études critiques à propos des textes, le concept d'*hapax legomena* (chose dite une fois) a été élaboré pour signaler la propriété attachée à une unité textuelle de constituer un exemple unique d'utilisation dans un corpus donné. Dans la pratique, les copistes et les commentateurs ont souvent noté cette propriété, jugée exceptionnelle, à propos d'unités textuelles remarquables du point de vue de leur forme. Dès le début des études quantitatives appliquées aux textes et avant que les dépouillements textométriques ne soient systématiquement confiés à des ordinateurs, les textométriciens ont noté que le phénomène de l'hapaxie, loin de constituer une propriété exceptionnelle pour certaines formes rares, constituait au contraire un phénomène massif pour tout texte écrit dans une langue naturelle. Depuis la description de la structure quantitative du vocabulaire opérée par G. K. Zipf (cf. [Zipf, 1936]) on sait au contraire que dans la plupart des corpus de textes écrits en langue naturelle, la propriété de n'apparaître qu'une seule fois dans un corpus est partagée par un très grand nombre de formes du texte.

propriété d'hapaxie est partagée par un très grand nombre de formes du texte. De ce fait, l'ensemble du texte se trouve parsemé de formes de fréquence 1 et tout fragment du texte en contient un certain nombre plus ou moins proportionnel à sa longueur. La surabondance de formes de fréquence 1 dans un fragment particulier constitue un souvent le signe que le fragment est le lieu de descriptions et d'énumérations de termes qui ne seront plus employés par la suite. A l'inverse, l'absence relative de ces formes est souvent le signe que le fragment contient des répétitions de segments de textes dupliqués dans le corpus.

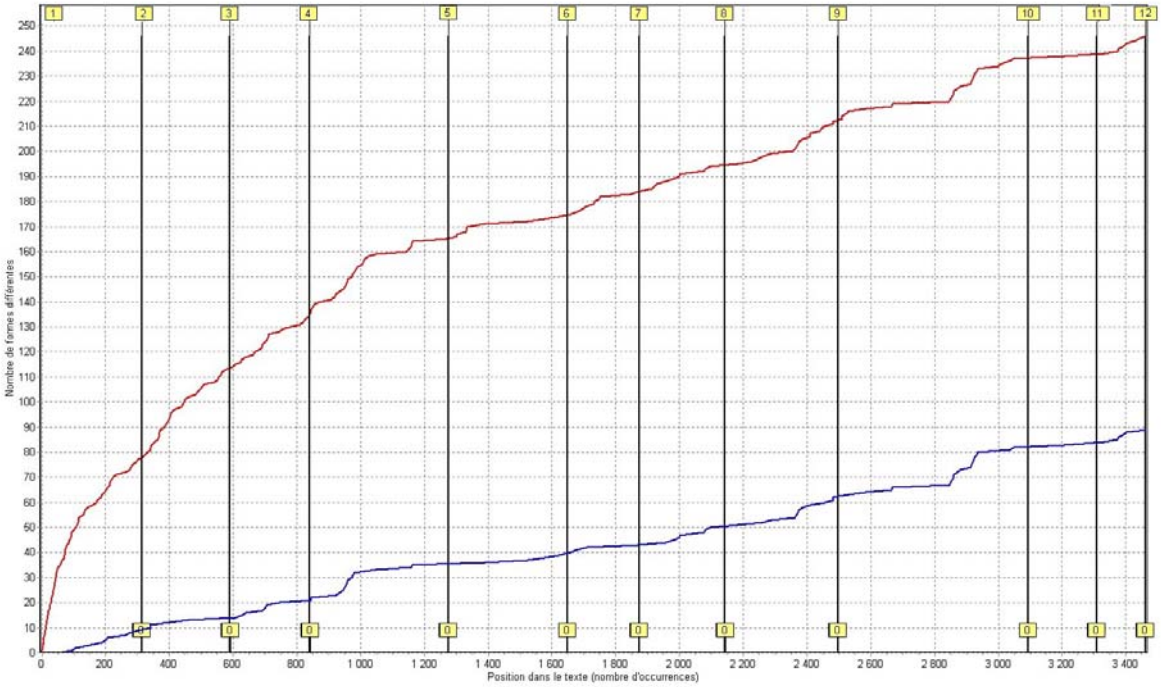


Figure : 5
Courbe d'accroissement du vocabulaire et courbe d'accroissement du nombre des hapax calculées pour le volet hiéroglyphique du corpus ***Naufragé***

==== Guide de lecture pour la figure 5 ====

- Le nombre des occurrences du texte se développe le long de l'axe horizontal entre le début et la fin du texte pour lequel la courbe a été établie.
- La *Courbe d'accroissement du vocabulaire* (en rouge, dans la partie supérieure du graphique) s'accroît d'une unité chaque fois que l'on rencontre une forme qui n'a pas encore été rencontrée précédemment. C'est une courbe croissante qui varie de 0 (au début du texte) à NbForm (nombre de formes différentes du texte, valeur atteinte lorsque le texte a été entièrement parcouru).
- La *Courbe d'accroissement du nombre des hapax* (en bleu, dans la partie inférieure du graphique) résulte d'un calcul similaire pour lequel ne sont prises en compte que les formes hapax du texte considéré (i.e. les formes qui ne possèdent qu'une seule occurrence dans l'ensemble du corpus). Cette seconde courbe varie de 0 à NbHap (nombre total des hapax du texte).

Dans le cas du dépouillement en signes hiéroglyphiques que nous avons adopté pour cette étude, l'unité de décompte concerne des unités dont les combinaisons permettent ensuite de former les unités plus étendues que sont les mots. Ces unités peuvent parfois coïncider avec des mots, dans d'autres cas elles n'en constituent qu'un élément. Compromis entre un

système basé sur un alphabet extrêmement réduit et un système dans lequel tous les signes auraient valeur d'idéogramme, le système d'écriture hiéroglyphique ne peut donc être totalement assimilé à un système lexical du point de vue de ses caractéristiques textométriques, ce dont témoignent d'ailleurs les décomptes produits au tableau 3.

Malgré ces différences, nous allons montrer que, la raréfaction des hapax constitue bien un signe de redondance du texte contenu dans le fragment par rapport à l'ensemble des fragments qui précèdent. Sur la figure 5, on peut vérifier que certaines portions du texte connaissent un accroissement faible du nombre des hapax (courbe d'apparition des hapax presque horizontale pour le fragment). La suite de notre étude nous permettra de vérifier que ces fragments constituent bien des reprises textuelles par rapport aux fragments précédemment rencontrés dans les parties précédentes du texte.

4.6 Étude des segments répétés du corpus

Les procédure de calcul des *segments répétés* permettent de localiser des suite de signes hiéroglyphiques apparaissant à l'identique à plusieurs endroits du corpus ***Naufragé***. Ainsi par exemple, la séquence de signes translittérés :

```
-A1 -r:f -n:V31A -mi -i -t*t:Y1 -i -r:y -xpr:r
```

peut être localisée à l'identique dans deux versets du texte (versets 21 et 125). L'éditeur du site *Projet Rosette* permet de rétablir la forme originale de cette séquence :

et de vérifier sa présence dans le texte original aux deux endroits indiqués[16]. On trouvera, figure 8, les traductions associées à cette séquence aux endroits du corpus qui la contiennent.

Classification et localisation des répétitions du corpus

Différents travaux consacrés à l'utilisation des recensements de segments répétés dans un corpus de textes montrent que les résultats fournis par ce type de formalisation renvoient la plupart du temps à des phénomènes textuels de niveaux très différents. Dans le cas des dépouillements en mots, les segments courts (i.e. composés de 2-3 formes) renvoient souvent à la présence d'unités lexicales complexes (mots composés, locutions, etc.) alors que la répétition de segments composés d'un plus grand nombre de formes trahit en général la présence de citations ou de reprises textuelles plus systématiques.[17]

L'analyse des segments répétés contenus dans chacun des volets du bitexte ***Naufragé*** fait apparaître toute une série de segments répétés particulièrement longs. L'établissement d'une concordance portant sur les segments les plus longs, tableau 5, nous permet de vérifier que plusieurs de ces segments trouvent une de leurs occurrences dans le fragment n°2 du conte à laquelle correspond une seconde occurrence qui peut être localisée dans le fragment n°5.

L'établissement d'une carte de sections sur laquelle on a signalé la présence des segments appartenant à ce seul groupe nous conduit au constat que la duplication de ces longues séquences résulte de la répétition d'un même récit, repris avec des variations à deux endroits différents du corpus (figure 4).

---

[16] Rappelons que l'identité que nous avons recherchée porte sur la *séquence* des signes élémentaires qui constituent la séquence hiéroglyphique. En l'occurrence, les deux versions de la séquence repérée présentent quelques écarts minimes qui peuvent concerner la disposition des signes sur la ligne.

[17] Sur la méthode des segments répétés, cf. par exemple [Salem 1994].

Tableau : 4

Extrait des concordances réalisées à partir des occurrences des segments répétés les plus longs dans le volet hiéroglyphique du corpus ***Naufragé***

Partie : 01Intro, Nombre de contextes : 1
- p : W - D : d - n : V31A - s - D : d - **A1** - r : f ! - n : V31A - mi - i - t * t :

Partie : 02VoyageEtNaufrage, Nombre de contextes : 8
: V31A - w - A1 - r - M14 - wr : r - S - **m** - d : p * t - P1 ! - n : t - mH : a - V1
: mD - mD : mD - m - s - x ! - w - iab - **s** - s - qd - d - A30 - A1 - V1 - V20 : V20
V1 - V20 : V20 - i - m - s ! - m - stp : **Y1** - n - km - m - t : niwt - mA : ir - A -
tA : N23 * Z1 - m - a : V31A - A - a ! - **ib** : Z1 - s - n : Z2 - r - mA : ir - A - w
r - S - tp : Z1 - a : Z1 ! - sAH - Y1 - **n** : 3 - tA : N23 * Z1 - f - A - t - A9 - a
m - i - i - t - A2 - n : U19 - nw - W - **i** - i - t - mw ! - i - m - f - n : t - mH
n - xt : t * Z1 - H - H ! - A25 - A24 - **n** - A1 - s - aHa - a : n - d : p * t - P1
a : n - d : p * t - P1 ! - m - t : Z6 - **n** : t - tyw - Z2 - i - m - s - D35 - z : p

Partie : 03IleDuKa, Nombre de contextes : 1
- H - n : a - A - p : d - w - zA : Z2 - **D35** : n - n : t * t ! - D35 : n - s - t - m

Partie : 04LeSerpent, Nombre de contextes : 6
V12 : Y1 - sw - w - r - xnt - n : t ! - **i** - w - wp : p - Z9 : n : f - r * Z1 : f -
: n - A1 - n - m : a - ini - n : t * w - **zp** : Z1 * Z1 - n : D - z : wr - A1 ! - n -
: n - iTi : t * t - A24 - i - m - A1 ! - **i** - w - wp : p - Z9 : n : f - r * Z1 : f -
- w - A1 ! - Hr - Z1 - X : t * Z1 - A1 - **m** - b - bA - A - H - D53 : Y1 - f ! - aHa
- n : A1 - n - m : a - ini - n : t * W - **zp** - Z1 * Z1 ! - n : D : z - wr - A1 - n -
- N36 - n : t * y - Aa13 : Z1 - f : y - **m** - n : U19 - nw - W ! - i - i - mw - aHa

Partie : 05RecitNaufAuSerp, Nombre de contextes : 9
- i - i - A1 - x - xA - A - m - D41 ! - **m** - b - bA - A - H - D53 : Y1 - f - D : d
p : p - w - t : D54 ! - sAq : sAq - G7 - **m** - d : p * t - P1 - n : t ! - mH : a - V1
- mD : mD - m - s - x : w ! - iab : Y1 - **s** - s - qd - d - A30 - A1 - V1 - V20 : V20
i - m - s ! - m - s - t : p - w - U21 : **Y1** - n : km - m - t : niwt ! - mA : ir - A
: N23 * Z1 ! - m - a : V31A - A - A24 - **ib** - Z1 - s - n : Z2 - r - mA : ir - A ! -
: N36 ! - tp - Z1 - a - Z1 - D61 - D54 - **n** : 3 - tA : N23 * Z1 - f - A - t - A9 - a
m - i - i - t - A2 - n : U19 - nw - w - **i** - i - t - mw ! - i - m - f - n : t - mH
- n - xt : t * Z1 - H - H - A19 - a ! - **n** : A1 - s - aHa - a : n - d : p * t - P1
n - d : p * t - P1 - m - t : Z6 : t ! - **n** : t - tyw - Z2 - i - m - s - D35 : z - p

Dans le cas de reprise textuelle d'un récit relativement long que nous venons d'explorer, on peut penser que l'existence d'une répétition n'aurait pas échappé à un lecteur attentif, pour peu que celui-ci soit suffisamment à l'aise avec la langue dans laquelle le texte a été rédigé. Une fois identifiées les zones de répétition, le repérage des unités textuelles qui n'apparaissent que dans l'un des deux fragments qui entrent en rapport de duplication peut alors permettre de localiser des variations entre les différentes versions du récit.

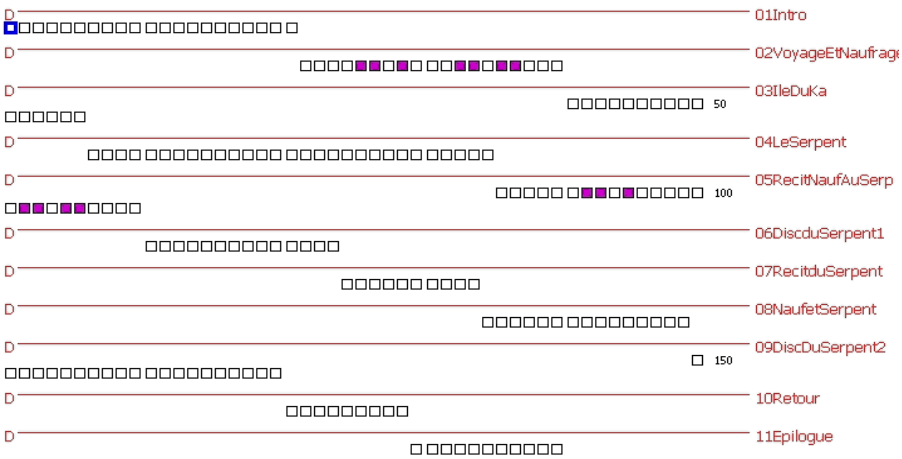


Figure : 6

Ventilation des occurrences des segments répétés longs trouvant dans les fragments 2 et 5 du volet hiéroglyphique du corpus ***Naufragé***

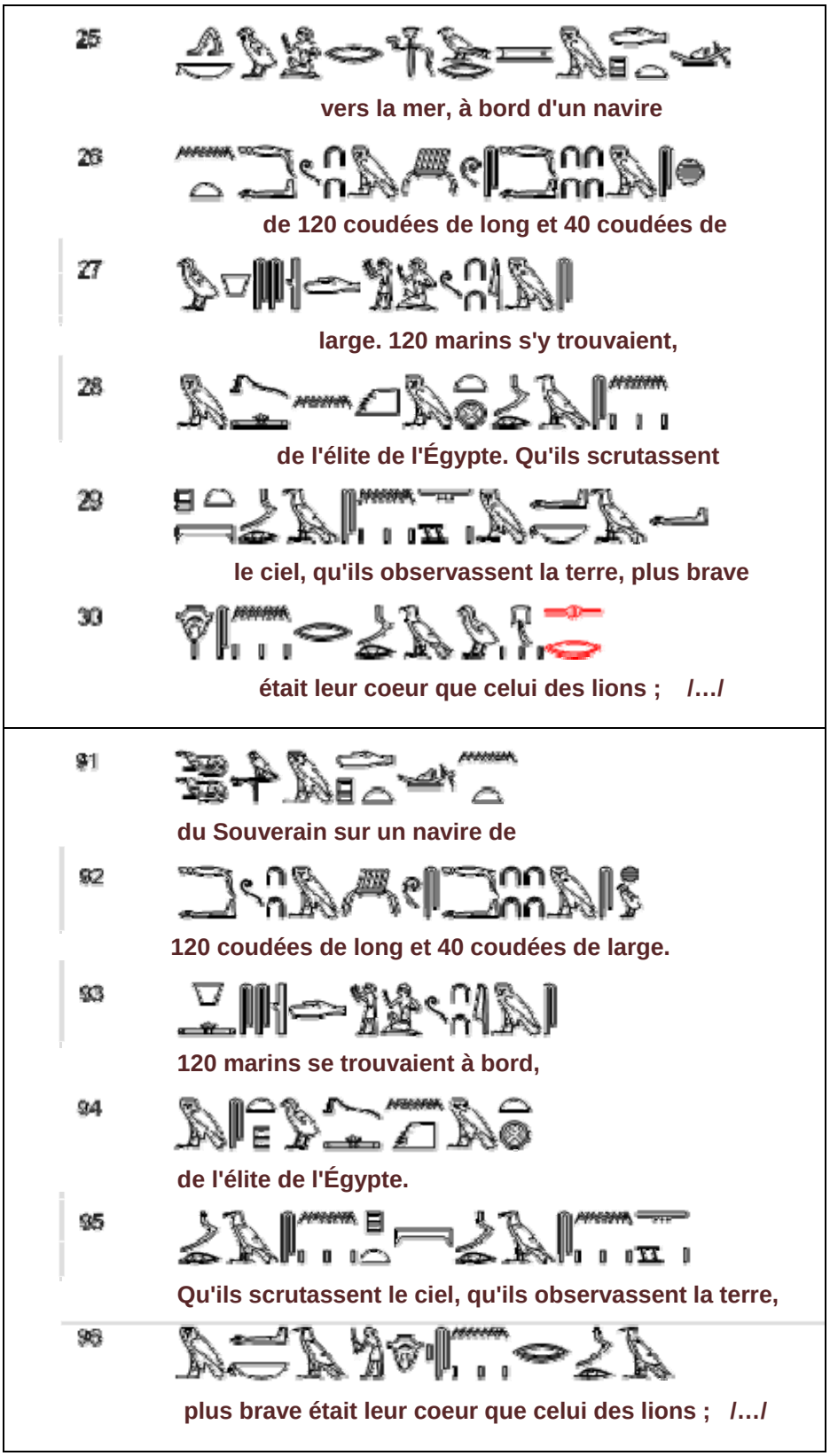


Figure : 7
Deux passages du corpus ***Naufragé*** rapprochés sur la base de leur utilisation de segments répétés communs**.**

La comparaison systématique entre les résultats fournis par la même méthode sur les deux volets du corpus multilingue peut permettre d'interroger utilement le travail du traducteur : a-t-il rendu par des formulations différentes des segments de texte absolument identiques dans le texte original ? a-t-il, au contraire traduit par les mêmes expressions des formulations qui différaient quelque peu dans ce même texte ?

récurrences isolées

La méthode des segments répétés permet également de repérer des récurrences moins systématiques dues à la reprise d'une formule particulière dont l'origine peut être trouvée soit dans l'existence d'un figement linguistique particulier soit au contraire dans la mise en pratique de procédés narratifs utilisés de manière récurrente. On voit par exemple sur la figure 8 le rapprochement que l'on peut opérer en suivant la même méthode entre les propos tenus par le vieux serviteur pour commencer le récit qu'il adresse à son supérieur et ceux prononcés par le Serpent pour commencer le sien.

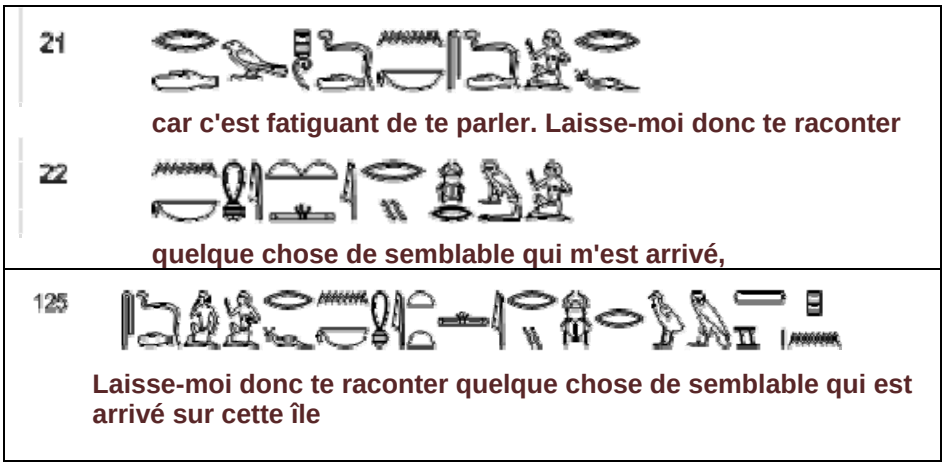


Figure : 8
Fragments du corpus ***Naufragé*** rapprochés sur la base de leur utilisation de segments répétés communs**.**

Dans ce second cas, la méthode textométrique apporte incontestablement un éclairage qui permet seul de localiser des répétitions segmentales importantes pour l'étude de la construction du récit, dans le cas du corpus que nous avons considéré et, a fortiori, dans le cas d'un corpus qui réunirait un plus grand nombre de textes.

5 Reproductibilité des explorations dans le bitexte

Dans ce qui précède, nous avons utilisé la traduction française du conte pour permettre au lecteur francophone de mieux s'approprier les résultats que nous obtenions à partir du volet hiéroglyphique du texte. Dans cette dernière section nous avons regroupé quelques résultats obtenus par la mise en œuvre des mêmes méthodes appliquées cette fois au volet français du bitexte. Ces résultats montrent que les phénomènes constatés sur le texte hiéroglyphique trouvent en quelque sorte un écho mesurable dans les résultats du même type que l'on obtient à partir de la traduction française.

Sur la courbe d'accroissement du vocabulaire établie à partir du volet français du corpus la stagnation est encore plus perceptible que sur la courbe réalisée à partir du volet

hiéroglyphique correspondant. Cette stagnation est encore plus marquée sur la courbe, située dans le bas du graphique, qui rend compte de l'apparition des hapax au fil du texte.

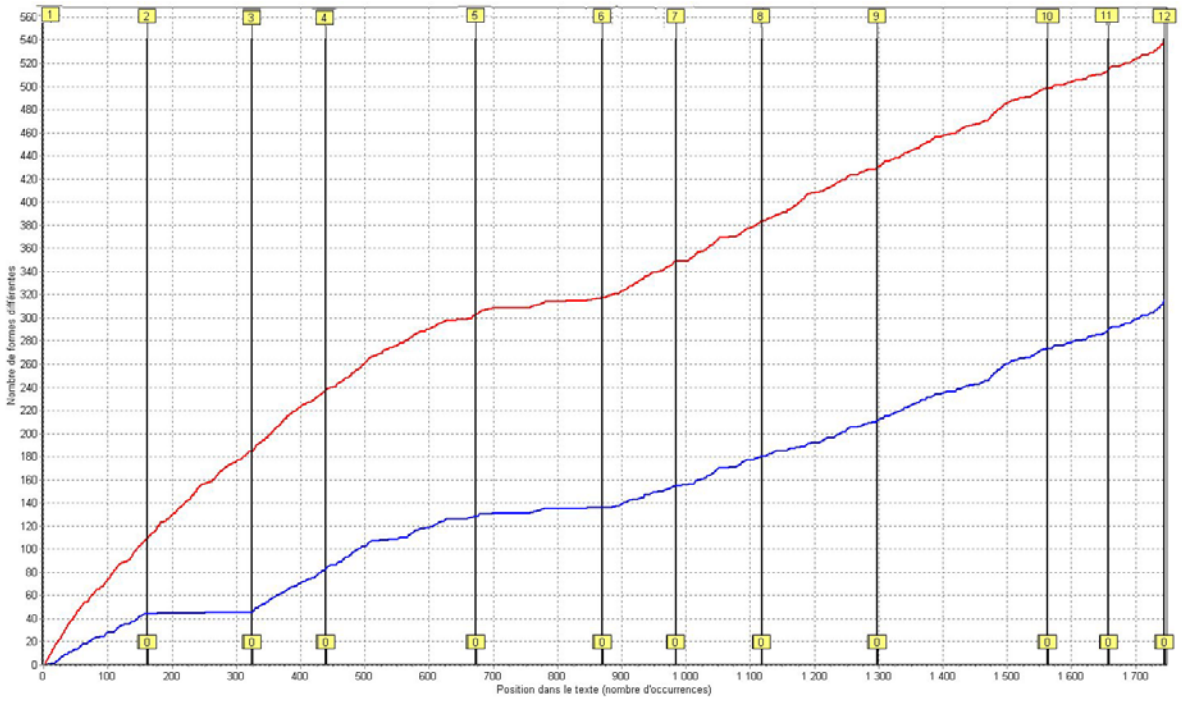


Figure : 9

Courbe d'accroissement du vocabulaire et courbe d'accroissement du nombre des hapax calculées pour le volet français du corpus ***Naufragé***

Comme c'était le cas pour le volet hiéroglyphique du corpus, la ventilation des segments répétés les plus longs montre une répartition privilégiée de certains segments entre les fragments 2 et 5 de la traduction française du conte.

On vérifiera sans surprise que les traductions françaises des deux parties constituées par la répétition d'un même récit dans le corpus original ont amené la création de textes qui sont très proches entre eux.

Tableau : 5

Extrait des concordances réalisées à partir des occurrences des segments répétés les plus longs dans le volet français du corpus ***Naufragé***

Partie : 01Intro, Nombre de contextes : 1,
§ car c ' est fatiguant de te parler . laisse - moi donc te raconter § quelque chose de
Partie : 02VoyageEtNaufrage, Nombre de contextes : 5
tais descendu § vers la mer , à bord d ' un navire § de 120 coudées de long et 40 coudées
large . 120 marins s ' y trouvaient , § de l ' élite de l ' égypte . qu ' ils scrutassent
a venue , un orage § avant son arrivée . une tempête était survenue § alors que nous
rvenue § alors que nous étions en mer et avant § que nous eussions touché terre . le vent
ta pas § un . et je fus déposé § sur une île par une vague de la mer . § je passai trois
Partie : 04LeSerpent, Nombre de contextes : 2
' il ouvrit la bouche vers moi , tandis que § j ' étais à plat ventre devant lui , §
. § il ouvrit sa bouche vers moi , alors que § j ' étais à plat ventre devant lui § "
Partie : 05RecitNaufAuSerp, Nombre de contextes : 5
les mines en mission § du souverain sur un navire de § 120 coudées de long et 40 coudées
. § 120 marins se trouvaient à bord , § de l ' élite de l ' égypte . § qu ' ils scrutassent
' y avait pas § de maladroit parmi eux . une tempête § était survenue alors que nous
urvenue alors que nous étions en mer , § avant que nous eussions touché terre . § " le
§ voici que j ' ai été déposé sur cette île par § une vague de la mer . § il me dit

6 Conclusion

Dans cette étude exploratoire portant sur un corpus de textes hiéroglyphiques, nous avons montré comment des méthodes textométriques pouvaient être requises pour explorer les répétitions segmentales à l'oeuvre dans un corpus de textes. L'étude de ces répétitions permet de mettre en évidence différents types de reprises textuelles : reprises de fragments étendus lorsqu'il s'agit de la répétition d'une portion de récit, reprises de fragments plus courts dans le cas de la répétition de formules, de locutions, d'expressions plus ou moins figées en langue.

L'étude d'un corpus de texte hiéroglyphique pratiquée en liaison avec celle de sa traduction alignée dans une langue plus accessible aux chercheurs contemporains (bitexte aligné) permet d'éclairer les résultats textométriques obtenus sur le corpus hiéroglyphique à l'aide de résultats du même type obtenus à partir de leur traduction. Cette possibilité permet d'envisager l'études systématique des traductions obtenues à partir de corpus hiéroglyphiques nettement plus vastes que le corpus réduit que nous avons considéré pour cette première étude.

 (S34 U28 S29) *Vie, prospérité, santé !*[18]

7 Références

Brunet, E., (2000). « Qui lemmatise, dilemme attise », in *Lexicometrica*, no 2.

Lamalle, C, Salem, A., (2002). « Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels », in Actes *des 6èmes Journées d'analyse des données textuelles,* St Malo.

Mayaffre, D. (2005). De la lexicométrie à la logométrie, *L'Astrolabe.*

Muller, Ch., (1963). « Le Mot, unité de texte et unité de lexique en statistique lexicologique », in *Travaux de linguistique et de littérature,* 1.

Salem, A. (1987). *Pratique des segments répétés*, Publications de l'INaLF, collection "St.Cloud", Klincksieck, Paris.

Zimina, M., (2004). *Approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles,* Thèse de doctorat , Université de la Sorbonne nouvelle – Paris 3, Paris.

Zipf, G. K., (1935). *The Psychobiology of Language, an Introduction to Dynamic Philology*, Houghton-Mifflin, Boston.

Webographie

Site du *Projet Rosette* : http://projetrosette.info/page.php?Id=1

Présentation et texte intégral du conte du naufragé :

http://pagesperso-orange.fr/sylvie.griffon/textes/naufrage/naufrage.htm

[18] Formule d'eulogie, (i.e.) courte proposition exclamative appelant toutes sortes de bénédictions sur la personne qui fait l'objet du texte, souvent placée à la fin des textes hiéroglyphiques égyptiens.