Revue, Informatique et Statistique dans les Sciences humaines
XXII, 1-4, 1986, pp. 179-190.

Constitution des fichiers-textes en néo-égyptien : lemmatisation et analyse automatique

J. WINAND

On ne peut douter que l'utilisation des moyens informatiques ouvre des voies nouvelles à la philologie et la linguistique égyptiennes. Il ne manque pas de domaines où l'application de ces méthodes peut se révéler particulièrement fructueuse pour notre discipline.

J'ai déjà eu l'occasion[1] de présenter un système d'encodage pour la constitution de fichiers de textes écrits en néo-égyptien. La saisie des informations s'effectue en transcription phonétique afin de permettre des opérations de type linguistique. Les recherches les plus récentes pour enregistrer les hiéroglyphes sur support informatique et les restituer montrent en effet qu'il est impossible, pour l'instant, de traiter automatiquement des phénomènes linguistiques à partir de données hiéroglyphiques. Cela est dû, pour une part, à l'ambiguïté du système graphique de l'égyptien et, d'autre part, aux difficultés techniques considérables pour mettre au point un logiciel suffisamment souple pour traiter des textes de nature ou d'époque différentes, où les codes graphiques ne se recouvrent pas exactement. Dans l'état actuel des choses, on peut espérer – et ce serait là un grand progrès – fusionner deux types de fichiers, le fichier texte et

1. J. Winand, *Analysis of Late Egyptian by Computer*, dans *Crossroad*, CNI Publications I (Copenhagen, 1986), pp. 389-400.

le fichier hiéroglyphique, de manière à permettre l'impression des résultats obtenus à partir du fichier texte et accompagnés des formes hiéroglyphiques correspondantes.

Il faut d'ailleurs reconnaître que le rôle du code graphique en tant que facteur discriminant a considérablement faibli à partir de la 20e dynastie dans les textes qui utilisent la langue vernaculaire. Dans l'égyptien de la première phase, il existe une sorte de convention orthographique qui fait que les mots sont presque toujours écrits de la même façon; il y a ainsi une adéquation étonnante entre le signifiant et le signifié.

Prenons, par exemple, deux thèmes différents mais qui offrent la même suite radicale *M + R*: mr_1, qui exprime l'idée de peine, de maladie et mr_2, qui exprime l'idée d'amour et de désir, représentés respectivement par les signes

 et

Bien que l'écriture égyptienne se soit formée par une sorte de rébus, il y a néanmoins des habitudes graphiques qui se sont créées et qui font que *mrw.t* "l'amour" ne sera jamais écrit

 .

On peut donc affirmer que ces tendances orthographiques procèdent, pour une part, d'une conscience instinctive de l'étymologie. En revanche, dès la fin du Nouvel Empire et, surtout, à la Troisième Période Intermédiaire, ce fragile équilibre a été battu en brèche afin de tenir compte de l'évolution phonétique. On trouve ainsi des graphies étonnantes rompant résolument avec la tradition (*cf. OAD*). Ce que je veux dire ici, c'est qu'en adoptant une transcription phonétique pour ce type de texte, on ne fait, somme toute, que se conformer à l'évolution du système graphique égyptien.

Dans les fichiers-textes, tels qu'ils sont réalisés à l'Université de Liège, chaque ligne d'enregistrement est réservée à une seule forme. Au stade final, une ligne comprend obligatoirement les éléments suivants :

1-18	lemme
19-38	forme
40-52	code du texte et références
54-65	analyse morpho-syntaxique
70	notes sur l'écriture
71	apparat critique

Les principes qui régissent ces différentes catégories ont été exposés ailleurs; je ne reprendrai donc que certains points de détail.

Le *lemme* constitue le mot tel qu'il figure dans le dictionnaire, c'est-à-dire, pour les substantifs, la forme du singulier; pour les adjectifs, la forme du masculin singulier et pour les verbes, la forme de l'infinitif.

La *forme* représente le mot tel qu'il se trouve dans le texte étudié. La solution qui a été adoptée ici est un moyen terme entre une transcription grammaticale stricte et une transcription de type grapho-phonétique, pour suppléer à l'absence des hiéroglyphes. Ceci permet de rendre compte de certaines nuances graphiques quand celles-ci ont une influence sur le niveau phonétique ou morpho-syntaxique. Par exemple, le substantif *s.t*, dans le syntagme circonstanciel *ḥr s.t=f* "sur son siège" pourra être noté de plusieurs manières différentes pour essayer de rendre compte des variations graphiques :

	Notation du lemme	Notation de la forme
	s.t	s.t
	s.t	s.ttw

Par contre, un graphème comme  ne sera pas noté autrement que dans le premier cas, car la variation n'affecte ici que le niveau grapho-sémantique.

On trouvera la liste des codes analytiques au tableau 1, ainsi qu'un extrait d'un fichier constitué (tableau 2).

Comme on peut le voir aisément, la réalisation de tels fichiers est relativement complexe. Elle demande beaucoup de minutie, de patience et de temps. Mais il faut aussi se souvenir que le fichier est constitué une fois pour toutes et que, passé cette étape, il devient la source unique de nombreuses exploitations.

Néanmoins, la lenteur de l'encodage restait un obstacle important. Avec cette méthode, un chercheur peut espérer encoder quelque 250 mots sur une journée, ce qui est évidemment très peu.

Après cette première étape obligée, je me suis préoccupé de la possibilité d'alléger la procédure en confiant à l'ordinateur les parties les plus contraignantes et les plus fastidieuses, c'est-à-dire la

Tableau 1
Liste des codes analytiques (extrait)

Catégorie grammaticale	Sous-catégorie	Genre Nombre Personne	Relation syntagmatique	Auxiliaire	Paradigme temporel	Relations syntaxiques	Emploi
Substantif 1	Nom commun 1 propre 2 composé 3	Masc. nM M / Fém. nM M Sg 1 4 A D Pl 2 5 B E Du 3 6 C F Invariable 0 Indéterminé X	Déb. du S 1 Mil. du S 2 Fin du S 3	Auxil. 1 Suit un A. 2 Suit un A. et est A. 3	A P sdm.f (pft) 1 A T sdm.f (prsp) 2 B t sdm.n.f 3 C emphat. 4 sdm.t.f 5 infin. 6 statif S partic. 8 G relative 7 impérat. 9 P I P séquent. p F III F F I f conjonc. K sdm.in.f I	indépend. 1 iw dép. D nty rel. R wn pas. P prép. conj. S séquent. C wnn temp. W rdi + F f. subst. I	Substantival 1 Adjectival 2 Adverbial 6
Adjectif 2	simple 1 nisbé subst. 2 nisbé prép. 3 article 4						
Numéral 3	cardinal 1 ordinal 2						
Adjectif-Pronom 4	suffixe 1 dépendant 2 indépendant 3 nouv. pron. 4 démonstratif 5 possessif 6 relatif 7 interrogatif 8 C.O.D. 9 conjonctif K	1 1 A D 2m 2 B E f 3 Z F 3m 4 C G f 5 P H neutre 6 invariable 0 indéterminé X					
Verbe 5	II III IV Fort 1 A G Fble 2 B H Gém. 3 C Réd. 4 D (-n) Caus. 5 E K C.gém 6 réd 7 F Irr. 8						
Adverbe 6	négation 1 lieu 2 temps 3 manière 4						
Préposition Conjonction 7	prép.simple 1 prép.comp. 2 conj.simple 3 conj.comp. 4						
Particule 8	proclitique 1 enclitique 2						

Tableau 2
Ounamon : 65,8 – 65,11

iw	iw	OUN0650804	5801 D	
iri	i.ir	OUN0650805	581214D 4	
i	i	OUN0650806	4112 1	
nw	nw	OUN0650807	520326	
r	r	OUN0650808	71 1 3	
p3	p3	OUN0650809	241 6	
kkwy	kkwy	OUN0650810	111 3	
r	r	OUN0650811	71 1 4 3I	
dd	dd	OUN0650812	5103 6 13I	
h3i	h3y	OUN0650813	5B4 21 4	
f	f	OUN0650814	414 1	
3tp	3tp	OUN0650901	5A1 2 4	
i	i	OUN0650902	411 1	
p3	p3	OUN0650903	241 6	
ntr	ntr	OUN0650904	111 2	
r	r	OUN0650905	71 4 3	
tm	tmy	OUN0650906	510 16 13C	
rdi	di.t	OUN0650907	580 2N C	
ptr	ptr	OUN0650908	5A4 2F 4	
sw	sw	OUN0650909	424 2	
ky	kti	OUN0650910	21A 6	
ir.t	ir.t	OUN0650911	11A 1	
iw	iw	OUN0651001	5801 C	
p3	p3	OUN0651002	241 6	
mr-mry.t	mr-mr	OUN0651003	1312 pC 1	3
iw	iy	OUN0651004	5803 6 4	
n	n	OUN0651005	71 3 3	
i	i	OUN0651006	411 3	
r	r	OUN0651007	71 3 4 3I	
dd	dd	OUN0651008	5103 6 13I	
smn	smn	OUN0651009	552 91 4	
tw	tw	OUN0651010	422 2	
r-ȝ3ˁ-r	ȝ3ˁ	OUN0651011	72 1 3	3
dw3w	dw3w	OUN0651012	111 3	
ḫr	ḫft	OUN0651101	584 1184	3
f	<f>	OUN0651102	414 1	

lemmatisation et l'analyse. Pour ce faire, je me suis inspiré de la méthode qui a été développée pour l'analyse des langues classiques, notamment du grec qui pose des problèmes de traitement fort semblables[1].

Je vais aborder maintenant les différentes étapes nécessaires à l'élaboration d'un fichier.

1. Constitution d'un thésaurus

La première tâche fut de constituer un thésaurus de formes à partir des fichiers qui avaient déjà été créés, c'est-à-dire établir un fichier contenant un exemple de toutes les formes qui avaient été lemmatisées et analysées. Ce travail fut effectué sur un corpus de 6 000 données représentant les *LRL* et le *Voyage d'Ounamon*. Ces deux fichiers ont été rassemblés puis triés suivant l'ordre logique de l'ordinateur et non d'après la suite alphabétique utilisée en égyptien. Cette manière de procéder accélère considérablement les opérations et n'est nullement gênante, puisque cette ordonnance ne sert que pour les opérations internes de la machine. Il va de soi que pour l'impression d'éventuels index ou concordances, l'ordre adopté est celui qui est traditionnel en égyptologie.

Après triage et élimination des formes semblables, un premier thésaurus de 1 500 formes différentes a été constitué (tableau 3).

2. Saisie des informations

Seule la forme telle qu'elle apparaît dans le texte est désormais encodée, toujours à raison d'une forme par enregistrement. Les codes de ponctuation nécessaires sont également indiqués dès cette première phase (tableau 4).

---

1. J. Denooz, *Le traitement des textes latins, grecs et français au Laboratoire d'Analyse Statistique des Langues Anciennes*, dans *Revista de la Universidad Complutense*, XXV, 102 (1976), pp. 143-167.

Tableau 3
Thesaurus (extrait)

ẖm`y.t	ẖm`y.t	11D 3	
ẖm`y.t	ẖm`y	11A 3	
ẖmsw	ẖmsw	1113 3	
ẖmsw	ẖmsy	111 3	
ẖmw	ẖmw	111 6	
ẖpsy	ẖpsy	211 6	
ẖpsy	ẖps	211 6	
ẖpsy	[ẖpsy]	211 6	2
ẖpsy	ẖp[sy]	211 6	2
ẖri	ẖri	111 B	
ẖri	ẖri	111 3	
ẖri	ẖri	210 13	
ẖri	ẖriw	215 6	
ẖsp	ẖp	5A03 6 4	
ẖsp	ẖsp	1A1 3	
ẖsp	ẖsp	5A 11 4	
ẖsr	ẖs	1A13 4	
ẖsr	ẖsr	111 3	
ẖsr	ẖsr	1113 4	
ẖw	ẖ3w3	5A0 26	
ẖ3i	ẖ3i.w	5B43 S 4	
ẖrdn	ẖ3rd3n3	111 3	
ḏḥwty-ḥtp	ḏḥwty-ḥtpw	1P1 A	
ḏḥwty-ms	ḏḥwty-ms	1P1 A	
ḏḥwty-ms	[ḏḥwty-ms]	1P1 A	
ḏḏ	ḏḏ	51 11 4	
ḏḏ	ḏḏ	510 6 13	
ḏḏ	ḏḏ	5102 6 1P	
ḏḏ	ḏḏ	51 11 4	
ḏḏ	ḏḏ	5102 6 13I	
ḏḏ	ḏḏ	512 91 4	
ḏḏ	ḏḏ	512 91 4R	
ḏḏ	i.ḏḏ	5102 6 1P C	
ḏḏ	i.ḏḏ	512 91 4	
ḏḏ	i.ḏḏ	5122 7 26R	

Tableau 4

iw		ptr	
i.ir		sw	
i		kti	
nw		ir.t	L
r		iw	
p3		p3	
kkwy		mr-mr	
r		iy	
dd		n	
h3y		i	
f	L	r	
3tp		dd	
i		smn	
p3		tw	
ntr		ȳ3ˋ	
r		dw3w	L
tmy		hft	
di.t		<f>	

3. Traitement des données

Tous les enregistrements du fichier sont alors numérotés; les formes sont ensuite triées suivant l'ordre logique de la machine. La numérotation servira, après le traitement, à retrouver l'ordre initial des mots dans le texte. On a donc, à ce stade, deux types de fichiers qui se présentent de manière fort semblable: un thésaurus de formes déjà lemmatisées et analysées, et un nouveau fichier qu'on désire traiter et dans lequel ne se trouvent que les formes. Dans les deux fichiers, rappelons-le, les formes se trouvent classées dans l'ordre logique de la machine. L'ordinateur a, dès lors, toutes facilités pour comparer les formes du fichier avec celles qui se trouvent dans le thésaurus. Lorsque se présente une forme qui figure déjà dans le thésaurus, l'ordinateur complète les parties manquantes, c'est-à-dire le lemme et l'analyse. Si plusieurs solutions sont possibles, elles sont toutes indiquées. Enfin, si le fichier contient des formes inconnues du thésaurus, elles sont simplement recopiées sans modifications.

Au terme de cette opération, le nouveau fichier ainsi constitué est à nouveau trié afin de retrouver sa séquence originale du

Tableau 5
Ounamon : 65,8

iw	iw	OUN0650804	5801
iw	iw	OUN0650804	5801 C
iw	iw	OUN0650804	5801 D
iw	iw	OUN0650804	5802
iw	iw	OUN0650804	5802 C
iw	iw	OUN0650804	5802 D
iw	iw	OUN0650804	58C3 S 4
iw	iw	OUN0650804	58C 2F 4
iri	i.ir	OUN0650805	58 1141 4
iri	i.ir	OUN0650805	582 91 4
iri	i.ir	OUN0650805	58 1 7 26R
iri	i.ir	OUN0650805	58 214D 4
iri	i.ir	OUN0650805	584 8 26
iri	i.ir	OUN0650805	58 2 7 12R
iri	i.ir	OUN0650805	58 2 7 13R
i	i	OUN0650806	411 2
i	i	OUN0650806	411 3
i	i	OUN0650806	411 5
i	i	OUN0650806	4112 1
i	i	OUN0650806	4112 5
i	i	OUN0650806	4112 pC 1
i	i	OUN0650806	4112 F1 1
i	i	OUN0650806	4112 PD 1
i	i	OUN0650806	4113 FS 1
nw	nw	OUN0650807	52 11 4
nw	nw	OUN0650807	52 21 4
nw	nw	OUN0650807	52 26
r	r	OUN0650808	71 1 3
r	r	OUN0650808	71 3 3
r	r	OUN0650808	73 3S
p3	p3	OUN0650809	241 6
kkwy	kkwy	OUN0650810	
r	r	OUN0650811	71 1 3
r	r	OUN0650811	71 3 3
r	r	OUN0650811	73 3S

texte (tableau 5). Il reste alors au chercheur, soit à éliminer les solutions inadéquates, là où plusieurs analyses ont été proposées pour une forme, soit à compléter la ligne d'enregistrement pour les formes qui n'ont pas reçu de solution satisfaisante (tableau 6).

Tableau 6
Ounamon : 65,8

iw	iw	OUN0650804	5801	
iw	iw	OUN0650804	5801 C	
iw	iw	OUN0650804	5801 D	—
iw	iw	OUN0650804	5802	
iw	iw	OUN0650804	5802 C	
iw	iw	OUN0650804	5802 D	
iw	iw	OUN0650804	58C3 S 4	
iw	iw	OUN0650804	58C 2F 4	
iri	i.ir	OUN0650805	58 1141 4	
iri	i.ir	OUN0650805	582 91 4	
iri	i.ir	OUN0650805	58 1 7 26R	
iri	i.ir	OUN0650805	58O214D 4	—①
iri	i.ir	OUN0650805	584 8 26	
iri	i.ir	OUN0650805	58 2 7 12R	
iri	i.ir	OUN0650805	58 2 7 13R	
i	i	OUN0650806	411 2	
i	i	OUN0650806	411 3	
i	i	OUN0650806	411 5	
i	i	OUN0650806	4112 1	—
i	i	OUN0650806	4112 5	
i	i	OUN0650806	4112 pC 1	
i	i	OUN0650806	4112 F1 1	
i	i	OUN0650806	4112 PD 1	
i	i	OUN0650806	4113 FS 1	
nw	nw	OUN0650807	52 11 4	
nw	nw	OUN0650807	52 21 4	
nw	nw	OUN0650807	52O26	—(03)
r	r	OUN0650808	71 1 3	—
r	r	OUN0650808	71 3 3	
r	r	OUN0650808	73 3S	
p3	p3	OUN0650809	241 6	—
kkwy	kkwy	OUN0650810	O O	—(111)③
r	r	OUN0650811	71 1 3	
r	r	OUN0650811	71 3 3	
r	r	OUN0650811	73 3S	

Ces nouvelles formes seront alors ajoutées au thésaurus qui s'enrichit ainsi à chaque nouveau traitement.

*

* *

De cette manière, il est possible d'encoder environ 3 000 formes par jour (à titre indicatif, le *Voyage d'Ounamon* ne dépasse pas 2 500 enregistrements). Toutefois, l'opération ultime, c'est-à-dire la sélection de l'analyse correcte parmi celles qui ont été proposées par la machine, réclamait encore beaucoup d'attention et donc beaucoup de temps. Il faut savoir que le volume du fichier, après lemmatisation, a été quintuplé. En effet, plus les codes morpho-syntaxiques sont précis, plus ils sont nombreux, c'est-à-dire qu'une forme, hors de tout contexte, est susceptible parfois d'une multitude d'analyses.

Aussi, il m'a semblé intéressant de faire lever par l'ordinateur un maximum d'ambiguïtés, une fois les formes traitées et remises dans l'ordre du texte. J'ai donc élaboré, en ce sens, un programme dans lequel figurent des règles syntaxiques du néo-égyptien permettant la résolution d'un certain nombre de cas par considération du contexte. Voici deux exemples destinés à mieux faire comprendre le système utilisé.

Reprenons d'abord le syntagme circonstanciel dont on a eu besoin précédemment, *ḥr s.t=f*, et supposons qu'il apparaisse tel quel dans un texte qu'on désire traiter. Après la lemmatisation et l'analyse automatiques, plusieurs solutions, notamment d'analyse, ont été proposées pour *s.t* et pour *-f*. *S.t* peut, en effet, figurer dans un complément d'objet direct, un complément circonstanciel, un génitif direct ou indirect, etc. De même, le pronom suffixe *-f* peut tout aussi bien être le sujet d'une forme verbale suffixale, l'élément pronominal d'une préformante, un génitif direct, le régime d'une préposition, etc. Or, il est clair que ce n'est que par la considération du contexte qu'on peut espérer lever ce type d'ambiguïté. Au cours de la procédure, l'ordinateur lit le fichier d'une manière séquentielle; quand il arrive à *ḥr*, il note qu'il est en présence d'une préposition grâce au code d'analyse (*cf. supra*). On suppose ici que l'ambiguïté entre le substantif *ḥr* "le visage" et la préposition *ḥr* a déjà été levée. L'ordinateur pourra ainsi effectuer un premier tri parmi les analyses qu'il trouvera pour le mot suivant. De même, en ce qui concerne le pronom suffixe *-f*, le fait que le mot immédiatement précédent est un substantif réduit considérablement les possibilités d'analyse.

Le deuxième exemple est plus complexe, car il s'agit d'analyser un groupe verbal. Soit la séquence *iw bn iw.f stp* (forme négative d'un futur III en position dépendante). Après traitement, plusieurs analyses ont été proposées pour les deux *iw* : morphème lié du futur III, *iw* dépendant, *iw* séquentiel, *iw* comme verbe plein, etc. A la différence du premier exemple où les règles à appliquer sont toujours progressives, c'est-à-dire que la levée d'ambiguïté d'une forme se fait au moyen d'un critère déjà défini, on a besoin ici aussi bien de règles progressives que régressives. En d'autres termes, il faut considérer tout le contexte d'une forme et plus seulement ce qui précède. En effet, c'est grâce à la présence de la négation *bn*, entre autres, que le premier *iw* est obligatoirement le *iw* dépendant, et c'est par le biais du deuxième *iw* que *bn* ne peut être le prédicat d'inexistence, mais bien le morphème négatif. Il faut donc que le programme passe d'abord en revue tout le système verbal afin d'y déceler les éléments qui serviront de critères de démarcation.

Au terme du traitement, l'ordinateur indique les solutions qui ne lui paraissent pas satisfaisantes. Il ne subsiste plus que la ou les formes (rarement plus de trois) qui peuvent encore être envisagées. Après l'application d'un deuxième programme de ce type destiné à l'affinage, plus de 75 % des ambiguïtés ont été levées.

Au stade actuel de mes recherches, la prise en charge d'un texte de longueur moyenne (environ 3 000 données), depuis l'entrée des données jusqu'à la résolution définitive des ambiguïtés, demande deux journées de travail.

La rentabilité et l'efficacité du système ont ainsi été multipliées par six. De cette manière, je pense pouvoir affirmer qu'est tombée la dernière barrière qui nous séparait encore de la constitution rapide de fichiers importants sans lesquels toute exploitation, pour alléchante qu'elle soit, n'a que peu de valeur et d'intérêt[1].

1. Actuellement, les textes suivants ont été encodés : *Voyage d'Ounamon, LRL, OAD, p Vandier, Lettre de el-Hibeh.*