TRAITEMENT AUTOMATIQUE DU LANGAGE NATUREL EN MOYEN ÉGYPTIEN

Serge Rosmorduc *

1 Introduction

Nous concevons actuellement, au LIFAC [1] un analyseur syntaxique du moyen égyptien. Il s'agit de construire automatiquement, à partir d'un texte (par exemple codé dans le format du « *manuel de codage* » [BGH+88]), une représentation de la structure syntaxique de celui-ci.

Si ce but n'est pas nouveau, l'approche envisagée s'inscrit néanmoins dans une optique relativement nouvelle. En effet, l'informatique linguistique des années 70 et 80 s'est surtout attachée à la construction de modèles détaillés de sous-ensembles de la langue. Cette façon d'aborder les choses conduit à privilégier le modèle par rapport aux sources, et à délaisser l'usage de corpus. Cependant, l'informatisation croissante des procédés d'édition a suscité un changement de point de vue, et un regain d'intérêt pour les corpus, et corollairement pour les méthodes statistiques.

Les textes qui sont actuellement disponibles sous forme informatique, qu'il s'agisse de livres ou de journaux, le sont en volume important. Ils sont souvent dégradés ; les coquilles sont nombreuses, et les marques diacritiques fréquemment oubliées ; il est ainsi parfois difficile de séparer un titre du texte qu'il précède. Pour traiter ces données, il n'est pas envisageable, du moins dans un avenir proche, d'espérer en faire une analyse très fine. L'approche classique du langage naturel, qui est de modéliser les « *phrases correctes* », échouerait, quand bien même, et ce n'est pas actuellement le cas, elle fonctionnerait parfaitement. Il faut donc décrire la langue

*LIFAC, ENS de CACHAN,
61 avenue du Président WILSON, 94235 CACHAN cedex
rosmord@lifac1.ens-cachan.fr,
http://lifac1.ens-cachan.fr/~rosmord

1. Laboratoire d'Informatique Fondamentale et Appliquée de Cachan

sans grande finesse, mais la décrire toute, en minimisant le nombre des erreurs, à défaut de pouvoir les éviter. Dans l'état actuel des connaissances, tout traitement automatique de la langue appelle un contrôle *a posteriori*. Le traitement a donc un intérêt si la vérification demande un travail moins important que l'analyse directe par un opérateur humain.

Cette approche, extrêmement pragmatique, convient à des corpus de textes qui ne sont pas destinés en premier lieu à l'analyse automatique ; elle nous semble donc convenir particulièrement bien au moyen égyptien, et aux langues mortes en général.

2 Nos objectifs

Le but de notre analyseur n'est en aucune façon la traduction automatique, mais, beaucoup plus prosaïquement, (et à long terme !) la simplification de la tâche du philologue lorsqu'il doit dépouiller de gros volumes de textes.

On peut discerner plusieurs usages pour une analyse automatique : ce peut être une aide interactive, suggérant des analyses possibles ; ce peut être aussi un outil d'exploration ; et enfin, elle peut fournir une analyse *a priori,* d'un un large corpus, afin d'accélérer la création de bases de données de textes.

2.1 Analyse exhaustive d'une phrase

Il peut arriver, face à un passage particulièrement obscur, que l'on souhaite disposer de la totalité des analyses syntaxiquement possible. La machine peut alors fournir une aide, qui ne sera utile que si les propositions du système sont en nombre suffisamment restreint pour être prises en compte par l'utilisateur. Cela impose de ne pas traiter tous les phénomènes syntaxiques : ainsi, l'attachement des circonstancielles introduit-il une grande ambiguïté, qu'il est impossible de lever dans le cas général en recourant à la seule syntaxe ; il en va souvent de même pour les datifs, difficilement distingables des génitifs.

2.2 Identification de constructions

Si l'on recherche dans un corpus toutes les occurrences d'une construction donnée (mettons, par exemple, que l'on s'intéresse à tous les sujets possibles du verbe ), la problématique est légèrement différente : la question posée guide les recherches, mais dans un texte de taille importante. Il ne s'agit plus forcément de donner la liste exhaustive des constructions possibles, mais de reconnaître les occurrences potentielles d'une construction particulière.

2.3 Aide à la construction de bases de données

La plupart des traitements automatiques effectués sur des textes fournissent des résultats imparfaits. Ils ont vocation à être complétés par un correcteur humain. Cependant, si la qualité générale de l'analyse est satisfaisante, une part importante du travail d'analyse du corpus est effectuée.

L'ordinateur a déjà été utilisé pour faciliter la lemmatisation de textes, par exemple par Jean WINAND dans [Win86].

Mais dans l'optique qui nous intéresse, l'ordinateur proposera *une* solution, éventuellement fausse, qui sera corrigée à la main.

3 Mise en œuvre

Les systèmes de traitement du langage naturel séparent généralement l'analyse lexicale et les traitements syntaxiques ; le nôtre ne fait pas exception à cette règle. Cependant, l'analyse lexicale se trouve, du fait du très grand nombre de variantes graphiques, particulièrement délicate.

L'analyse syntaxique, quant à elle, se trouve confrontée au problème de l'ambiguïté des constructions. Ce n'est une surprise pour personne. Néanmoins, si ce problème se pose pour l'analyse de *toutes* les langues, les connaissances que nous avons du moyen égyptien sont souvent proches de celles qui sont formalisables et peuvent donc être fournies à un ordinateur ; une phrase contenant un ou deux hapax conduit généralement à une multitude de traductions et d'interprétations, les problèmes qui se posent au traducteur humain recoupant alors partiellement ceux qui se posent à la machine.

3.1 Analyse morphologique

Nous allons, dans cette partie, détailler les diverses opérations nécessaires à l'analyse morphologique. Le texte, en entrée, sera *a priori* dans le format défini par le *Manuel de codage.* Il importe de décider si les symboles grammaticaux que celui-ci comporte seront utilisés ou non. Dans un premier temps, nous avons choisi la simplicité en supposant que nous disposions au moins de la séparation des mots, afin de pouvoir travailler sur l'analyse syntaxique. Néanmoins, nous donnerons une idée de traitement susceptible de séparer les mots.

Les traitements détaillés ci-dessous peuvent fournir, classées par ordre de préférence, plusieurs translittération possibles pour un même mot.

3.1.1 Séparation des mots d'un texte

On peut, en disposant d'un lexique correct, dresser la liste des séquences observables de 2 signes. Une séquence absente du lexique est probablement une fin de mot ; ainsi dans la figure 1. Pour ne pas introduire de fins de mots abusives, il faudra travailler sur un lexique comportant des formes fléchies, en particulier les pluriels.

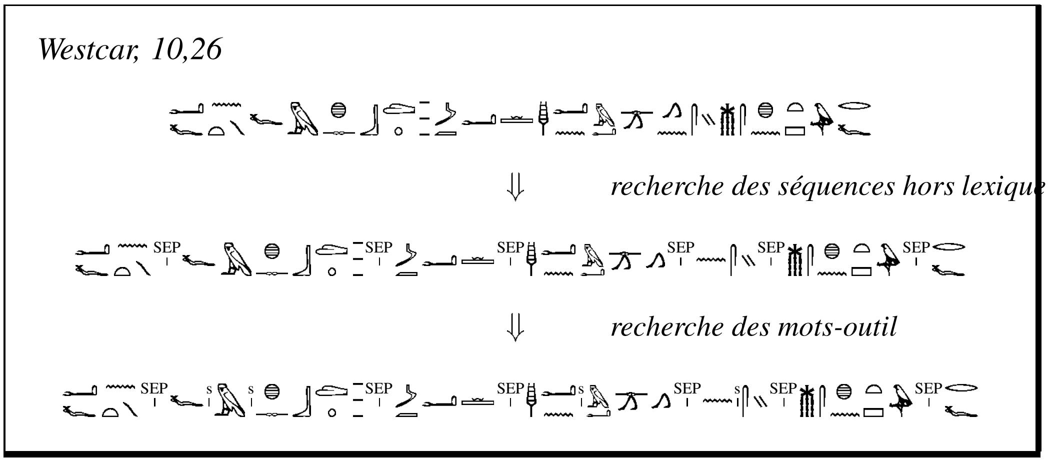

FIG. 1 - *Découpage en mots*

Il faut ensuite traiter le texte ainsi coupé, en prêtant une attention particulière à la recherche des mots fréquents, qui ne se terminent pas forcément par un déterminatif ; par exemple, la seconde séquence de la figure 1, , commence par deux mots-outils, le pronom et la préposition . Comme les mots de ce type sont en nombre limité, il est concevable d'en dresser une liste quasi exhaustive.

Nous envisageons aussi d'utiliser des méthodes statistiques, en particulier celles qui ont été utilisées, avec succès, en lemmatisation automatique. Le principe que nous appliquerons est le suivant : la transition entre deux signes successifs est dotée de trois valeurs possibles ; elle peut être interne à un mot, séparer deux mots, ou séparer un mot d'un suffixe (au sens large du terme).

Ensuite, sur un corpus déjà renseigné, l'on calcule les probabilités qu'un signe soit placé dans un environnement donné, c'est à dire : $P(-S-)$, $P(=S-)$, $P(/S-)$, $P(-S=)$, $P(=S=)$, $P(/S=)$, $P(-S/)$, $P(=S/)$, $P(/S/)$ où S représente le hiéroglyphe, $-$ représente la liaison entre deux signes dans un même mot, / la liaison entre deux mots, et = la séparation entre un mot et une terminaison grammaticale.

Ainsi, notre texte de la figure 1, complètement traité, donne-t-il : / - - -

```
5 regle L1 *** A/[X] *** L2
==>  L1 *** A/[(X,1)] *** L2.

20 regle L1 *** C/[X,Y] *** A/[X] *** B/[Y] *** L2
==> L1 *** C/[(X,1),(Y,2)] *** A/[(X,1)] ***
     B/[(Y,2)] *** L2.

100 regle 'D4'/[i,r] *** D21/[r]
==> 'D4'/[(i,1),(r,2)] *** D21/[(r,3)]
```

FIG. 2 - *Règles de simplification*

Un corpus ainsi traité nous permet d'estimer les probabilités liées à un signe donné. Le court exemple que nous donnons fournit ainsi des renseignements sur le comportement du signe : on trouve deux fois = / et une fois - -. Disposant, pour chaque signe, des probabilités citées plus haut, on pourra, face à un texte inconnu, calculer le découpage le plus probable. Pour limiter la taille des calculs, on peut utiliser le traitement évoqué au début.

3.1.2 Translittération

Pour pouvoir translittérer un mot, il faut pouvoir décrire la façon dont les signes se combinent ; préciser, par exemple, que, si se lit *ʿ*, si se lit *ꜣ*, et , *ʿꜣ*, l'ensemble se lit, non *ʿꜣʿꜣ*, mais *ʿꜣ*.

La méthode employée pour translittérer les mots est la suivante : nous définissons des règles de simplification, qui permettent de décrire la façon dont des signes se combinent.

La première règle de la figure 2 signifie qu'un unilitaire peut se lire tout seul. Par exemple, dans le mot , le se lit *s*, sans être en relation avec aucun autre signe (alors que dans , il se combine avec pour donner la lecture *ms*). `L1` et `L2` désignent le début et la fin du mot, `A` est le hiéroglyphe, et `[X]` est une de ses translittérations, composée d'une unique lettre.

La seconde règle précise qu'un bilitère suivi de deux unilitaires correspondant à sa lecture doit se lire seul (par exemple :). Dans cette règle, `A` est un unilitaire qui peut se translittérer `X`, et `C` un bilitère qui peut se translittérer `[X,Y]`. Dans le cas qui nous intéresse, `A` est , `X` est *ʿ* ...

La seconde partie des règles (après le symbole ==>) permet de préciser quelles consonnes sont effectivement présentes, et dans quel ordre. Les nombres qui interviennent par exemple dans `[(X,1)]`, permettent, de spécifier, d'une part, que la consonne identifiée à `X` se lit avant celle identifiée à `Y`, et que les deux consonnes `X` qui interviennent ne doivent être lues qu'une fois.

La règle a une priorité, qui permet d'ordonner l'application des simplifications ; ainsi, la troisième règle, dont la priorité est 100, sera-t-elle appliquée pour le mot , entraînant la lecture *ỉrr,* au lieu de la lecture *ỉr* qu'aurait entraînée l'application de la seconde règle.

Selon toute probabilité, la première solution trouvée par notre système ne sera pas toujours la bonne. Aussi celui-ci peut-il revenir sur ses décisions et proposer plusieurs lectures.

3.1.3 Recherche dans le lexique

Considérons une graphie dans un texte, par exemple . L'algorithme de translittération nous fournira plusieurs valeurs possibles : *wꜥrt, wꜥrtỉnr...* les déterminatifs seront repérés comme tels.

Le problème se posera alors de retrouver le mot dans notre lexique, sachant que les graphies enregistrées ne correspondent pas forcément à celle dont nous disposons. La translittération nous permettra de limiter le nombre de mots examinés. Pour distinguer les homophones, nous utiliserons, d'une part les déterminatifs, et, d'autre part, les signes bilitères ou trilitères composant les mots.

Nous associons à chaque déterminatif une série de traits sémantiques. ainsi, possède-t-il le trait « terrain ». Les traits peuvent être plus ou moins précis, et certains en impliquent d'autres : par exemple, les idéogrammes de dieux particuliers impliquent le trait « dieu », et le trait « oiseau » est partagé par les signes   ...

Supposons donc que notre lexique contienne deux mots se translittérant *wꜥrt :* , « *plateau désertique* », et ,« jambe ».

Les signes ayant une valeur de déterminatifs sont ici : , et . Le premier est le déterminatif du « gebel », le second possède deux interprétations : « jambes » ou phonogramme ; le troisième est le déterminatif de la « pierre ».

La structuration des traits, donnée dans la figure3, permet de conclure que « gebel » et « pierre » étant plus proches l'un de l'autre que de « jambe », c'est le sens « plateau désertique » qui sera choisi.

Un tel système oblige à structurer plus fortement son lexique qu'il n'est d'usage. Le mot ,« district », se place ainsi naturellement sous la même entrée que « plateau désertique ».

Quand les déterminatifs ne suffisent pas à distinguer les homonymes, il est possible de recourir à la comparaison des trilitères puis des bilitères qui forment le

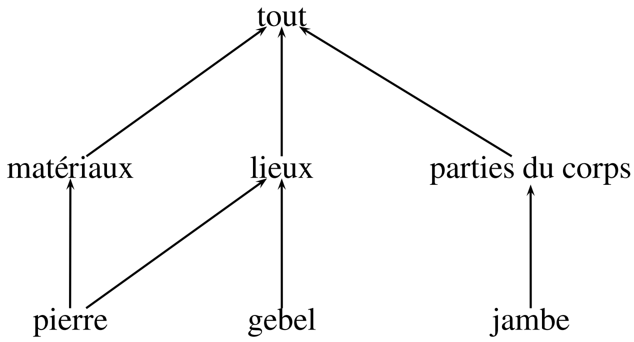

FIG. 3 - *Structure des traits*

mot.

Enfin, les abréviations, les graphies « sportives », et toutes les autres formes qui sortent par trop des règles, se trouvent dans le lexique à la place où l'algorithme de translittération les range, ce qui permet de renvoyer commodément à leur vraie valeur : ainsi, est-il enregistré à *nw,* et renvoie-t-il à *m-ḫnw* (c'est d'ailleurs la méthode employée par les dictionnaires). De même, sera-t-il enregistré sous *nṯrsnṯr.*

3.2 Analyse syntaxique

Les méthodes mises en œuvre sont relativement classiques, mais visent, au contraire de ce qui est généralement développé, à la robustesse de l'analyse. Aucune phrase trouvée dans notre corpus ne doit être rejetée comme incorrecte, mais l'analyse doit pouvoir rester partielle. C'est possible en ne poussant pas l'analyse très loin, dans un premier temps, pour ensuite la raffiner. Cette méthode est employée dans [Con91] pour réaliser un correcteur grammatical, et est discutée dans [Eje93]. L'idée est que certains problèmes de structures, ne trouvant pas de solution syntaxique simple, peuvent engendrer énormément d'ambiguïté. Ainsi, la distinction entre la préposition et le nisbé *ny.* Déterminer que dans le groupe : il faut comprendre, selon la figure 4, « *alors la majesté du roi Khéops dit* ... » et non, selon la figure 5, « *alors la majesté dit au roi Khéops ...* » peut à la rigueur être réalisé par l'analyseur lexical, quand celui-ci gérera les expressions figées, et donc en particulier la construction « *la majesté du roi N* » pour désigner un monarque.

Malheureusement, dans une phrase telle que « *alors le roturier dit à la femme d'Oubainer ...* », l'interprétation de *n* comme un datif plutôt que comme un génitif est difficile à réaliser automatiquement. Il sera de même ardu de décider de la forme d'un verbe ; cependant, s'il a un

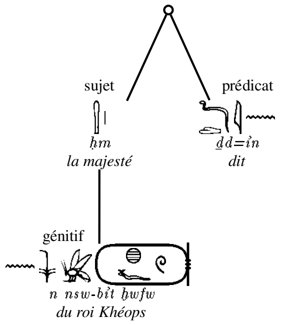

FIG. 4 - *Analyse*

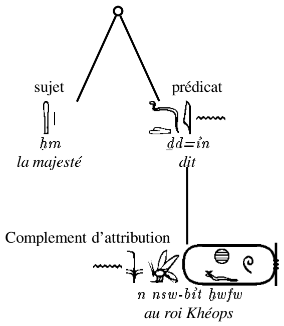

FIG. 5 - *Analyse*

sujet, le lien entre le sujet et le verbe existera que ce dernier soit un prospectif, une forme verbale relative, ou quoique ce soit d'autre.

Une première analyse regroupe donc des constructions sûres, sans forcément lever toutes les ambiguïtés qui s'y trouvent. Une seconde série de traitements aura pour tâche de préciser l'analyse et d'éliminer les mauvaises hypothèses. L'ampleur de cette seconde analyse peut énormément varier, selon le but de l'analyse : un système peut servir d'aide à la traduction, par exemple en proposant toutes les analyses grammaticales possibles d'une phrase ; il peut aussi servir à traiter un texte en vue de l'extraction d'information, par exemple une recherche grammaticale ou lexicale ; ou encore on peut demander au système de proposer une analyse, éventuellement corrigée par un spécialiste, en vue de l'établissement d'une base de donnée linguistique.

L'approche variera selon les cas. On pourrait espérer conserver une même description du langage, et simplement modifier la stratégie d'analyse. Si cette idée est séduisante, sa mise en pratique se heurte à des problèmes qui dépassent de loin le simple cadre technique.

4 Conclusion

Notre système est encore très réduit, et ses diverses composantes ne sont pas encore réunies (en particulier, la translittération et l'analyse syntaxique ne communiquent pas encore).

Il sera probablement nécessaire, pour lui permettre de traiter des textes d'origines diverses (par exemple, pour translittérer des hiéroglyphes aussi bien que de l'hiératique), de disposer de plusieurs bases de règles. Nous espérons que le cœur des dites bases ne variera pas, et que seule changeront les règles concernant la stratégie d'analyse.

Néanmoins, ce qui est déjà réalisé, nous incite à penser que le Moyen Égyptien n'est pas pire qu'une autre langue quant à la possibilité d'automatiser partiellement son analyse, et que cette automatisation peut se révéler, à terme, fructueuse.

Références

[BGH+88] Jan Buurman, Nicolas Grimal, Michael Hainsworth, Jochen Hallof, et Dirk Van Der Plas. *Inventaire des signes hieroglyphiques en vue de leur saisie informatique*. Mémoires de l'Académie des Inscriptions et Belles Lettres. Institut de France, Paris, 1988.

[Con91] Patrick Constant. « *Analyse syntaxique par couches* ». thèse de doctorat, École Nationale Supérieure des Télécommunications, 1991.

[Eje93] Eva EJERHED. « Nouveaux courants en analyse syntaxique ». *Traitement Automatique des Langues*, 34(1):61–82, 1993.

[Win86] Jean WINAND. « Constitution de fichiers-textes en néo-egyptien: lemmatisation et levée d'ambiguïté automatique ». *Revue. Informatique et statistique dans les sciences humaines*, 22:179–190, 1986.