La lemmatisation de l'arabe non classique

Laurence Tuerlinckx

U.C.L. – Institut orientaliste – Centre d'Études sur Grégoire de Nazianze
Place Blaise Pascal 1 – 1348 Louvain-la-Neuve – Belgique
tuerlinckx@ori.ucl.ac.be

Abstract

The current tools for the lemmatization and automatic analysis of the Arab language were conceived for classical or standard Arabic (literary corpora, newspapers...). Philological research on ancient texts written in non classical language (Christian texts in Middle Arabic in our case) requires a system which takes into account the graphic and syntactic variations of this language level. A program of formal lemmatization, adapted to these characteristics, was elaborate with the aim of creating lemmatized concordances and indexes. This software is based not on automatic analysis, but on recognition of forms already encountered, which makes it possible to gather the various forms under the same classical lemma, whatever their actualization in the texts. The lemmata are tagged with regard to their morpho-syntactic class as defined by Arab grammarians, and are classified by root, according to the model of traditional dictionaries. The philological options and methodological thoughts behind the treatment of this language (characterized by agglutination and non-vocalization) will be presented, as well as some prospects about exploiting this new data base.

Résumé

Les outils actuels de lemmatisation et d'analyse automatique de la langue arabe ont été conçus pour l'arabe classique ou standard (corpus littéraires, journaux…). Les recherches philologiques portant sur des textes anciens rédigés en une langue non classique (textes chrétiens en moyen arabe dans notre cas) nécessitent donc un système qui prenne en compte les variations graphiques et syntaxiques de cet état de langue. Un programme de lemmatisation formelle, adapté à ces particularités, a été élaboré dans le but premier de créer des concordances et index lemmatisés. Ce logiciel, basé non sur l'analyse automatique, mais sur la reconnaissance des formes déjà rencontrées, permet de classer les formes sous un même lemme classique, quelle que soit leur actualisation dans les textes. Les lemmes sont étiquetés relativement à leur classe morphosyntaxique telle que définie par les grammairiens arabes et sont regroupés par racine, suivant le modèle des dictionnaires classiques. Les options philologiques et la réflexion méthodologique concernant le traitement de cette langue (caractérisée par l'agglutination et la non vocalisation) seront présentées, ainsi que quelques perspectives d'exploitation de cette nouvelle base de données.

Mots-clés : arabe, lemmatisation, forme, lemme, racine, analyse, morphologie, philologie

1. Introduction

Depuis plusieurs décennies déjà, des recherches sont poursuivies dans le cadre du traitement automatique de la langue arabe. L'un des premiers théoriciens de ce domaine, David Cohen propose un essai d'analyse automatique dès 1961 (Cohen, 1961/1970). Partant d'une analyse morphologique minimaliste, basée sur le principe que toute forme linguistique arabe se traduit en schème et racine, les recherches vont se développer depuis les premiers travaux sur le lexique et la morphologie jusqu'à la mise au point d'analyseurs automatiques, de systèmes d'indexation, de correcteurs, etc. De nombreux projets sont en cours et il existe des bases de données disponibles proposant des corpus divers sous forme électronique, traités automati-

quement[1]. Un traitement automatique de l'arabe est donc non seulement possible, mais réalisé et en cours de perfectionnement (notamment dans le domaine de la traduction automatique).

Cependant, dans le cadre bien précis de nos recherches, nous avons été amenés à créer notre propre programme de lemmatisation. Après un exposé du problème, nous présenterons ce logiciel, ainsi que la réflexion méthodologique qui a présidé à son élaboration. Les premiers résultats obtenus et surtout les développements envisagés seront ensuite décrits.

2. La lemmatisation des textes en arabe non classique

Les outils disponibles pour le traitement automatique de l'arabe, conçus pour la langue classique, ne sont pas adaptés aux recherches philologiques basées sur des textes anciens rédigés en arabe non classique.

2.1. *Problème des états de langue en arabe*

Dès les premières réflexions sur le traitement automatique de l'arabe, le champ d'investigation a été défini : « *Par ARABE il est entendu ici L'ARABE LITTÉRAIRE, langue écrite de la littérature et de la presse, et parlée ordinairement à la radio, dans les cours et conférences universitaires et scientifiques, dans les discours officiels, etc., dans tous les pays arabes. Il devra être également admis, bien que cela n'ait aucune incidence sur la théorie, qu'il s'agit de L'ARABE MODERNE, ce qui permettra dans les essais éventuels de négliger des faits qui peuvent caractériser l'usage de la littérature la plus ancienne.* » (Cohen, 1970 : 49)[2]. Cette limitation du domaine pris en compte, justifiée du point de vue méthodologique, est liée à la question particulière des divers états de langue. L'arabe, langue sacrée du Coran, connaît une grande stabilité dans un créneau bien précis qui est celui de la littérature classique et des milieux de l'enseignement, et aujourd'hui, le domaine de la culture officielle, de la presse, etc. C'est l'arabe standard ou littéraire, universellement partagé par les lettrés de tous les pays arabes. Par contre, parallèlement à cette lignée relativement figée si ce n'est l'élargissement du lexique, il existe de nombreuses branches s'écartant plus ou moins de la norme. L'arabe dialectal dans toutes ses variétés, essentiellement oral, et le moyen arabe (état intermédiaire entre le dialectal et le classique) essentiellement écrit, sont autant de réalisations différentes d'une même source. Suffisamment proches pour constituer une seule et même langue, suffisamment éloignées pour ne pas s'intégrer dans les systèmes de traitement automatique, ces variétés linguistiques demandent une autre approche.

2.2. *La lemmatisation de l'arabe*

Sur le plan théorique, les principes généraux de la lemmatisation s'appliquent également à la langue arabe et les définitions établies dans le cadre du « Projet de recherche en lexicologie grecque », nous ont servi de postulat de départ : « *Au niveau du lexique de la langue, les* lemmes *servent d'adresses lexicales aux* formes*; les* formes *sont, au niveau du discours, les actualisations des* lemmes*. Classer les* formes *sous les* lemmes *adéquats, et définir les* lemmes *de manière pertinente, contraint le "lemmatiseur" décrivant le vocabulaire d'un auteur*

[1] Par exemple: The An-Nahar Newspaper Text Corpus, proposé sur le site de l'ELDA (http://www.elda.fr) ou l'analyseur morphologique mis au point par Ken Beesley : the Xerox Arabic Morphological Analysis and Generation, disponible en version demo sur le site du Xerox Research Centre Europe : http://www.xrce.xerox.com/competencies/content-analysis/arabic/ Le dictionnaire électronique Sakhr : http://www.aramedia.com/diction.htm. Nous ne savons pas ce que sont devenus le projet DIINAR-MBC et son analyseur AraParse (Dichy, 2001 : 1 ; Ouersighni, 2001 : 1).

[2] Cette limite du champ d'investigation des recherches est unanimement partagée par tous ceux qui travaillent sur le traitement automatique de l'arabe. Voir par exemple les titres des conférences dans ACL/EACL, 2001.

et (re)construisant progressivement le lexique d'une langue, à porter un regard sur cette langue et à l'analyser »[3]. Si les principes développés en fonction du grec ancien valent pour l'arabe, il y a cependant certaines particularités de cette langue qui demandent une adaptation dans la manière de concevoir la lemmatisation.

2.2.1. *La lemmatisation de l'arabe standard ou classique*

Outre l'orientation droite-gauche, les principales caractéristiques de la langue arabe, langue sémitique, par opposition aux langues indo-européennes sont l'agglutination, la structure particulière combinant schème et radical, et la non vocalisation.

Le *mot* ou l'*unité graphique*, « *suite de graphèmes entre deux blancs* » (ou signes de ponctuation) correspond le plus souvent en arabe non pas à une *forme* ou « *unité susceptible de figurer sous une entrée lexicale ou lemme* »[4], mais à une suite de formes collées les unes aux autres. Les mots du texte sont des *formes agglutinées* (ex. conjonction + préposition + article + nom) ; théoriquement cela nécessite une étape préliminaire de séparation des formes, mais concrètement, c'est la forme graphique entière qui est lemmatisée, c'est-à-dire ramenée à une succession de lemmes distincts.

La structure du mot arabe est donc décomposable en cinq éléments : proclitique, préfixe, base, suffixe et enclitique. La *base* est une combinaison de lettres radicales (le plus souvent trois) et d'un schème (redoublement, augment, voyelles longues et brèves, etc.). La base – avec préfixe et suffixe – forme le *noyau lexical*, éventuellement entouré d'extensions (Dichy, 1997 : 296-297 et 2001 : 3-4). Les analyseurs automatiques travaillent à partir des combinaisons valides de ces éléments.

Les voyelles brèves ne sont pas indiquées dans les textes courants. En règle générale, seuls le Coran et les textes à vocation didactique sont vocalisés. Cette caractéristique entraîne un haut degré d'ambiguïté, les formes non vocalisées peuvent représenter un nombre élevé d'actualisations différentes. Le traitement automatique de l'arabe doit pouvoir traiter des textes vocalisés comme des textes non vocalisés.

2.2.2. *La lemmatisation de l'arabe non classique : l'exemple du projet « Grégoire de Nazianze »*

Le projet « Grégoire de Nazianze » a pour objet l'ensemble des versions orientales d'une même œuvre : les discours de Grégoire de Nazianze, Père de l'Église et auteur grec byzantin (4^{e} siècle)[5]. La version arabe de ces discours, datable du 10^{e} siècle, témoigne de cet état de langue appelé « moyen arabe » (voir 2.1.) ; les éditions (Grand'Henry, 1988 et 1996 ; Tuerlinckx, 2000) ne corrigent pas le texte, ni ne le standardisent. Le texte édité, à lemmatiser ne s'inscrit donc pas dans le registre classique. Les textes rédigés en moyen arabe présentent des particularités graphiques, morphologiques et syntaxiques qui s'écartent de l'arabe classique (Blau, 1966). Les modifications graphiques ont pour conséquence immédiate que l'analyseur conçu pour l'arabe classique ne reconnaît pas la forme entrée. Certaines lettres faibles faisant partie de la racine (', *w*, *y*) reçoivent un traitement différent en moyen arabe (particulièrement dans la conjugaison) ce qui empêche la reconnaissance des formes. L'ordre

[3] Ce projet est développé à l'Institut orientaliste de l'Université catholique de Louvain, à Louvain-la-Neuve. Cette citation est extraite de la présentation du projet et du *Thesaurus Patrum Graecorum* sur le site http ://tpg.fltr.ucl.ac.be. Voir aussi Coulie, 2003.

[4] Définitions de P. Tombeur, citées dans Coulie (1996 : 37).

5 Site du projet : http://nazianzos.fltr.ucl.ac.be

des mots fixé par des règles précises en arabe classique, n'est pas respecté et une simplification de la syntaxe (disparition des modes, modification des genres, etc.) rend invalides certaines grammaires sous-jacentes à l'analyse automatique et à la levée d'ambiguïté. De plus le lexique particulier des textes chrétiens comporte un nombre important et illimité de termes d'origine non arabe, ne rentrant donc pas dans la structure de base vue plus haut[6].

Jusqu'à présent, les éditions de cette version arabe ne comportent pas d'index, car seul un index lemmatisé permet des recherches sur le lexique et la langue. Un index des formes du texte est facilement réalisable, mais sans intérêt étant donné le nombre important d'éléments adventices placés devant la base.

L'approche multilingue du projet vise à la comparaison des versions latine, arménienne, syriaque, arabe et géorgienne ; cette option nécessite un même système de référence au texte grec (langue source) dans toutes les langues traitées. Le système de référence doit être compatible avec celui de la concordance lemmatisée de ces textes dans le *T.P.G.* (Mossay et Cetedoc, 1990), en vue de la réalisation de lexiques multilingues et d'alignement des contextes.

3. Le programme de lemmatisation « GLOR.ARABE »

Afin de répondre aux besoins spécifiques mentionnés ci-dessus, un logiciel de lemmatisation a été créé grâce aux efforts conjoints d'arabisants et d'informaticiens[7].

Ce logiciel ne s'appuie pas sur des traitements automatiques de l'arabe développés ailleurs, et ceci en raison d'un choix, motivé d'une part par les exigences précises que nous posons en tant que philologues, d'autre part par les conclusions négatives obtenues à l'issue d'une enquête concernant l'accessibilité aux outils informatiques existants (contacts restés sans réponse, manque de suivi dans les projets exposés, incompatibilité avec le projet multilingue de notre équipe et surtout champs et objectifs trop éloignés des nôtres).

3.1. *Description de la méthode et des caractéristiques techniques*

Le programme GLOR.ARABE résulte de la mise au point de deux applications : l'une concerne les prétraitements, l'autre consiste en un outil interactif d'analyse linguistique élémentaire. Le langage de programmation utilisé est *MS Visual Basic* et les procédures et formulaires sont réalisés sous *MS Access*.

3.1.1. *Prétraitements*

Les opérations de prétraitement comportent quatre points :

– la découpe du texte en unités graphiques ou « formes agglutinées », accompagnées d'une série de références (auteur, œuvre, livre, chapitre, paragraphe, ligne, nœud ou numéro d'ordre du mot) ;

– la réorganisation du texte en vue d'une version finale (après ajouts, suppressions de mots, corrections, modification des sauts de lignes) ;

[6] Le même phénomène existe bien sûr en arabe classique ; Dichy appelle « pro-bases » ces noms issus d'emprunt. (Dichy, 1997 : 296).

[7] Les principaux concepteurs de ce logiciel sont Boris Maroutaeff, informaticien de l'équipe Informatique Facultaire, et Aurélie Gribomont, doctorante à l'Institut orientaliste ; les principes théoriques ont été discutés et établis par J. Grand'Henry et L. Tuerlinckx (tous sont membres de la Faculté de Philosophie et Lettres de l'Université catholique de Louvain à Louvain-la-Neuve). Ce logiciel a été baptisé GLOR.ARABE du nom du Département d'Études Grecques, Latines et Orientales de cette faculté.

– la réalisation d'unités de contextes ou concordances (5 mots avant le mot-clé, 5 mots après) ;

– l'établissement d'index et de tables en vue de la publication d'un index lemmatisé des textes.

3.1.2. *Analyse linguistique*

L'analyse linguistique développe trois procédures :

– l'encodage d'entrées de dictionnaires ou « lemmes » pour les différentes formes du texte (création d'un *Thesaurus* arabe) ;

– la lemmatisation (première) automatique et interactive au fur et à mesure de l'enrichissement du *Thesaurus* des lemmes (reconnaissance de formes et de lemmes) ;

– une procédure de navigation à l'intérieur du texte et du *Thesaurus* en vue de la proposition d'analyse.

3.1.3. *Caractéristiques techniques*

Ces applications requièrent le système d'exploitation *MS Windows XP Pro* et le traitement de texte *MS Office XP Pro*. Les données textuelles sont du format Unicode (police Arial unicode MS arabe) ; contrairement à ce qui se fait habituellement en traitement automatique de l'arabe, nous travaillons directement sur caractères arabes et non sur transcription.

Le texte peut provenir d'un fichier *Mac* (*Nisus* par exemple) ; il doit alors être converti en fichier *Windows* à l'aide du programme *Mac2Win*.

3.2. *Le formulaire de lemmatisation*

Le texte encodé doit être nettoyé de toute ponctuation, lignes blanches, etc., les sauts de ligne doivent être marqués (dans le cas d'un texte déjà édité, il faut respecter les lignes de l'édition). Le texte ainsi préparé est converti en un fichier '.txt' qui est rattaché à la base de donnée sous un numéro d'identification. La lemmatisation peut alors commencer. Après sélection de l'œuvre à traiter dans la liste des textes encodés, un formulaire de lemmatisation s'affiche, avec une série de champs remplis (contexte, références, cadre de propositions de lemmes) ou à remplir (cadre de lemmatisation).

La ligne supérieure situe le terme à lemmatiser dans son *contexte*. La *forme agglutinée* est le « mot graphique » composé de une ou plusieurs formes. Les références sont celles de la division interne à l'œuvre, le *nœud* étant le numéro d'ordre du mot graphique dans le texte. Ce numéro est définitivement attribué au mot ; les additions ou suppressions de mots suite à une correction d'éditeur par exemple, se traitent par le biais du *nœud caché*. Ce système de référence est parfaitement compatible avec le système de référence utilisé par les lemmatiseurs du *T.P.G*. Le cadre *lemmatisation* comportera autant de lignes qu'il y a de formes dans la forme agglutinée, chaque forme reçoit un *lemme*, auquel sont attachées une *nature* et une *racine*. Le *numéro de lemme* indique la position de la forme en question à l'intérieur de la forme agglutinée ; ce numéro n'est donc pas lié au lemme lui-même, mais à sa forme actualisée. Enfin, la case *emprunt* est cochée lorsque le lemme n'est pas un mot arabe, mais une transcription ou une adaptation d'un mot étranger. Le cadre *propositions* fournit, le cas échéant, des propositions de lemme pour la ou les différentes formes composant la forme agglutinée.

3.3. *Principe de la reconnaissance des formes*

Le principe de base de ce programme est non pas l'analyse, mais la reconnaissance de formes déjà rencontrées, aussi aberrantes et peu classique soient-elles. La reconnaissance fonctionne à deux niveaux différents : au niveau de la forme agglutinée et au niveau du lemme de chaque forme séparée.

Face à une forme du texte, le logiciel propose les différentes combinaisons de lemmes ou les différents lemmes qui ont déjà été attribués à cette même forme dans un quelconque texte lemmatisé. Il suffit alors de choisir le(s) lemme(s) qui convien(nen)t et de le(s) rattacher au mot traité.

Le second niveau de reconnaissance, intervient dans la lemmatisation de la ou des forme(s) d'un mot graphique. Lorsqu'on commence à encoder le lemme d'une forme non reconnue, un menu déroulant propose les lemmes déjà enregistrés et une recherche peut se faire sur le lemme entier ou sur les premières lettres (recherche partielle). Le résultat fournit un ou plusieurs lemme(s), accompagné(s) de leur nature, de leur racine, et de un ou plusieurs numéro(s) d'ordre. (Ex. l'article proposé en 1re, 2e ou 3e position).

Si aucun lemme proposé ne convient, il faut encoder un nouveau lemme et lui adjoindre ses éléments distinctifs (numéro, nature, racine, emprunt). L'actualisation de ces données entraîne la reconnaissance du nouveau lemme lorsqu'une forme identique sera rencontrée.

4. Aspects théoriques de la lemmatisation arabe

Les champs à remplir dans le formulaire de lemmatisation sont autant d'éléments à déterminer qui, ensemble, définissent le lemme. Chacune de ces déterminations s'appuie sur des choix théoriques et pose des problèmes spécifiques à résoudre.

4.1. *Lemme – Nature – Racine : définitions*

La lemmatisation réduit les formes à leur entrée de dictionnaire ; en arabe comme dans les autres langues, on se conforme à l'usage des dictionnaires classiques[8].

4.1.1. *Lemme*

Le lemme est l'entrée de dictionnaire, il s'agit donc de la forme classique et entièrement vocalisée à laquelle se rattache la forme du texte (classique ou non, vocalisée ou non) :

– les *verbes* sont ramenés à la 3e pers. sg. de l'accompli actif, sauf dans le cas de certains verbes figés n'ayant qu'une conjugaison partielle ;

– les *noms variables* sont ramenés à la forme du nominatif sg. (masc. pour les noms qualificatifs), les *noms invariables* à leur forme classique vocalisée (masc. sg. pour les pronoms) ;

– les *particules* sont ramenées à leur forme classique vocalisée.

À la différence d'autres langues comme le grec, on ne peut dire qu'il y a en arabe des «*formes indéclinables et indéclinées dans un texte, qui sont donc leur propre "forme de départ" ou "entrée de dictionnaire"* » (Coulie, 1996 : 38). La non vocalisation et la variation orthographique de la langue (voir 4.2.2.) demandent une lemmatisation distincte de la forme.

Les lemmes composés sont gardés comme composés si le composé lui-même fait l'objet d'une entrée dans les dictionnaires classiques. La nature qui lui est attribuée est celle du com-

[8] Notre principal dictionnaire de référence est celui de la Ligue arabe : ALECSO (1989).

posé, la racine retenue est celle du premier élément, afin d'assurer au composé sa juste place dans l'ordre alphabétique.

4.1.2. *Nature*

À chaque lemme est associée une étiquette précisant sa nature morphosyntaxique. En arabe, les catégories de mots se réduisent à trois : les verbes, les noms et les particules. Certains grammairiens ajoutent une catégorie « instrument » recoupant plus ou moins celle des particules ; nous avons limité cette catégorie à l'article, afin de le distinguer des autres particules. L'analyse suivie ici est celle des grammairiens arabes[9], non celle des arabisants occidentaux, auteurs de grammaires anglicisées ou francisées. Ainsi, par exemple, il n'y a pas de catégorie « adverbe » ou « préposition » et les participes ne sont pas rattachés à la catégorie « verbe », mais à la catégorie « nom ». Ces catégories se subdivisent en notions plus précises : 85 natures différentes ont été distinguées.

En dehors de l'analyse automatique, cet étiquetage morphosyntaxique est utile pour plusieurs raisons :

– il permet de lever des ambiguïtés entre certains lemmes et de distinguer certains lemmes homographes ;

– il intervient dans la structure du dictionnaire, puisque sous les racines, les lemmes sont classés selon une préséance due à leur nature et non selon l'ordre alphabétique ;

– il fournit un élément intéressant dans l'étude du lexique d'un texte.

4.1.3. *Racine*

Le plus souvent trilitère, la racine est la suite des consonnes formant le radical du mot ; à chaque racine correspond un champ sémantique. La racine est un élément important dans les langues sémitiques : associée à un schème, elle forme la base du mot (voir 2.2.1.). Les lemmes sont tous rattachés à une racine ; ce sont les racines qui sont classées alphabétiquement dans le dictionnaire, non les lemmes (voir 4.1.2.).

Racine est à prendre ici au sens large : il s'agit, sur le plan pratique, de fournir une référence qui détermine la place du lemme dans le dictionnaire. Dans certains cas, sous « racine » se trouve une suite de lettres qui ne constitue pas un radical, mais reprend le lemme dépourvu de ses voyelles brèves et d'éventuels éléments adventices : ce sont les *lemmes-racines* (noms propres, emprunts, transcriptions, onomatopées, particules, etc.). La case « emprunt » cochée est souvent associée à un lemme-racine.

La racine permet également de distinguer des lemmes homographes : deux combinaisons de lettres radicales et de schèmes différents peuvent aboutir au même résultat, alors qu'il s'agit de deux lemmes distincts.

4.2. *Problèmes particuliers*

4.2.1. *Homographie et spécifications*

Des spécifications sont ajoutées aux lemmes, non seulement pour fournir une information complète sur les entrées du dictionnaire, mais aussi pour réduire les cas d'homographie. Ces spécifications consistent à ajouter

[9] La grammaire suivie est en grande partie celle de Chartouni (Grand'Henry, 2000).

– à tous les verbes de la 1re forme: la forme conjuguée à l'inaccompli. Par contre, la voyelle de l'accompli est précisée dans la nature (V1a, V1i, V1u), parce que cette voyelle régit l'ordre des verbes 1 dans le dictionnaire. Des voyelles de conjugaison différentes peuvent s'accompagner de sens distincts et donc justifier des lemmes distincts.

– à certains noms singuliers: le(s) pluriel(s) brisé(s) (irréguliers). Des pluriels différents peuvent prendre des significations nettement distinctes et dans ce cas, ils justifient la création de lemmes distincts.

– à certaines particules : une notification précisant le type de particule (notification ajoutée à la nature) ou son usage (notification ajoutée au lemme). Des usages nettement différents d'une même particule justifient la création de lemmes distincts.

4.2.2. *Variations orthographiques*

Les variations orthographiques sont répandues, même à l'intérieur d'un même texte. L'éditeur d'un texte ancien peut rendre dans son édition la variété présente dans les manuscrits, si ses principes d'édition le justifient. Cela ne pose pas de problème pour la lemmatisation, puisque le lemme suit toujours l'orthographe classique. Cependant une variation peut subsister d'un dictionnaire à l'autre, particulièrement au niveau de la vocalisation ; en ce cas, les différentes vocalisations sont données sous un même lemme. Le dédoublement du lemme ne se justifie que si la variation vocalique s'accompagne d'une modification de sens.

4.2.3. *Ambiguïté*

L'absence de voyelles brèves dans les textes et l'écriture sporadique de signes tels que le redoublement ou l'allongement, entraînent une très forte ambiguïté. Généralement, cette ambiguïté est facilement résolue par l'association de formes, le sens, le contexte, etc. Dans certains cas, peu fréquents, rien ne justifie le choix d'un lemme plutôt qu'un autre, les différents lemmes possibles sont alors mentionnés, mais avec le même numéro d'ordre dans la forme, indiquant ainsi qu'ils occupent la même position.

4.2.4. *Référence*

Les noms propres et leurs dérivés, ainsi que les emprunts à d'autres langues posent un problème de référence. Ces formes présentent une grande variété orthographique dans les textes et il n'existe dans le domaine arabe, ni ouvrage de référence absolue, ni bases de données suffisamment développées pour pouvoir utiliser un critère tel que la fréquence d'un usage ou son ancienneté, comme cela peut se faire en grec par exemple.

5. Résultats obtenus et développements

Le travail de lemmatisation n'en est qu'à ses débuts, mais un dictionnaire commence à se former et les requêtes permettent d'effectuer certaines recherches sur le lexique des textes étudiés.

5.1. *Résultats – État d'avancement*

Deux discours de Grégoire de Nazianze ont été lemmatisés (Discours 1 et 27), deux autres sont en cours de lemmatisation (Discours 45 et 40) ; 24684 formes séparées ont reçu un lemme, ce qui correspond à 2692 lemmes différents encodés.

En quelques chiffres[10] :

	Discours 1	***Discours 27***	***Discours 45***	***Discours 40***
Formes agglutinées	825	2173	(2650 / 6865)	(7250 / 11181)
Formes séparées	1297	3407	(3988)	(15121)
Verbes	152 (11,7%)	378 (11%)	427 (10,7%)	1888 (12,4%)
Noms	611 (47,1%)	1558 (45,7%)	1858 (46,5%)	6679 (44,1%)
Particules	371 (28,6%)	1001 (29,3%)	1222 (30,6%)	4818 (31,8%)
Instruments/ articles	160 (12,3%)	467 (13,7%)	479 (12%)	1738 (11,4%)

On constate la stabilité des rapports entre les différentes catégories de lemmes, les pourcentages sont proches d'un texte à l'autre. Le rapport entre formes agglutinées et formes séparées varie peu : en moyenne chaque forme du texte se compose de 1,5 forme. (Plus dans le Discours 40, mais la lemmatisation en cours est ici sélective et ne reflète pas la moyenne).

Les requêtes et états en *Access* permettront de fournir un index lemmatisé des textes édités et en cours d'édition, et de mener dès à présent certaines recherches approfondies sur le lexique et les caractéristiques de cette version arabe. (Par ex. la fréquence élevée de telle catégorie morphosyntaxique comme résultat d'un procédé de traduction bien précis).

5.2. *Développements*

Cette nouvelle base de donnée demande à être enrichie de textes arabes issus d'époques et de milieux divers, musulmans et chrétiens, classiques et non classiques. Cela permettrait de développer des études synchroniques et diachroniques du lexique arabe.

Nous avons grandement bénéficié de l'expérience des auteurs du *T.P.G.* (*Thesaurus Patrum Graecorum*) et du *D.A.G.* (Dictionnaire automatique grec)[11] ; ces travaux nous ont servi de modèles et ont nourri notre réflexion sur la lemmatisation. La compatibilité de notre *D.A.A.* (Dictionnaire automatique arabe) naissant, avec le *D.A.G.* rend possible des développements multilingues. Le lien de traduction unissant une ou plusieurs forme(s) du texte grec à une ou plusieurs forme(s) du texte arabe, sera automatiquement transformé en un lien de lemme(s) à lemme(s), et le système commun de référence au texte facilitera la correspondance entre le texte traduit et sa source. Les développements récents du *D.A.G.* en un *D.D.G.* (Dictionnaire dérivationnel du grec) qui permet une interrogation à partir des morphèmes grecs (Gérard et Kindt, 2004) sont particulièrement intéressants pour nos recherches : le traducteur arabe avait tendance à décomposer le mot grec en morphèmes de la même manière que le fait le *D.D.G.*, avant de le faire passer dans la langue arabe, souvent élément par élément. Le lexique bilingue de nos textes devra se construire sur le lien de morphème(s) à lemme(s) plus que de lemme(s) à lemme(s) et l'exploitation du *D.D.G.* s'avère dès à présent une aide précieuse pour la compréhension des techniques de traduction (du grec à l'arabe) mises en œuvre dans les textes que nous étudions.

Le logiciel GLOR.ARABE lui-même est appelé à de nouveaux développements, car les syriacisants de notre équipe de recherche (CEGN) sont intéressés par une lemmatisation de ce

[10] État de la base de données en décembre 2003.

[11] Présentation du *D.A.G.* et de ses développements sur le site du *T.P.G.* : http://tpg.fltr.ucl.ac.be/PAGE.HTM.

type. L'adaptation du programme au syriaque est en cours d'élaboration ; le syriaque est une langue proche de l'arabe, appartenant au même groupe des langues sémitiques et partageant donc un certain nombre de caractéristiques communes. Cependant, des modifications sont nécessaires notamment en raison de l'imperfection du tri alphabétique syriaque en *Access*. De plus, certaines options théoriques sont différentes en ce qui concerne l'analyse morphologique ou les spécifications à ajouter aux lemmes.

Ces développements constituent un pas de plus dans la réalisation des projets multilingues de notre équipe.

Références

ACL/EACL (2001) *Workshop, Arabic Language processing: status and Prospects.* Toulouse 6 juillet 2001, http://www.elsnet.org/acl2001-arabic.html.

ALECSO (1989). *Al-mu'jam al-'arabiyy al-'asâsiyy*. ALECSO – Larousse.

Blau J. (1966). *A Grammar of Christian Arabic. Based mainly on south Palestinian texts from the first millennium. Corpus Scriptorum Christianorum Orientalium*, vol. (267 : Subsidia 27).

Grand'Henry J. (2000). *Grammaire arabe à l'usage des Arabes. Traduction française et commentaires des Éléments d'arabe, morphologie et syntaxe, II de Rachid Chartouni (Beyrouth).* Série Pédagogique de l'Institut Linguistique de Louvain, vol. (24). Peeters.

Cohen D. (1961/1970). Cohen D (1961). Essai d'une analyse automatique de l'arabe. *La Traduction Automatique*, vol. (2/3) : 48-71. Republié dans Cohen D. (1970). *Études de linguistique sémitique et arabe.* Mouton : 49-78. Les citations renvoient aux pages de cette réédition.

Coulie B. (1996). La lemmatisation des textes grec et byzantins : une approche particulière de la langue et des auteurs. *Byzantion*, vol. (66) : 35-45. Article disponible on-line à partir de la page : http://tpg.fltr.ucl.ac.be/PAGE.HTM

Coulie B. (2003). Corpus Christianorum. Thesaurus Patrum Graecorum. In Leemans J. (Ed.), *Corpus Christianorum 1953-2003. Xenium Natalicum. Fifty Years of Scholarly Editing*. Brepols : 169-172.

Dichy J. (1997). Pour une lexicomatique de l'arabe : l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot. *Meta*, vol. (XLII, 2) : 291-306.

Dichy J. (2001). On lemmatization in arabic. A formal definition of the Arabic entries of multilingual lexical databases. In *Proceedings of ACL/EACL*.

Gérard R. et Kindt B. (2004). D'un dictionnaire de lemmatisation (*D.A.G.*) à un dictionnaire dérivationnel du grec ancien (*D.D.G.*). In *Actes des JADT 2004*.

Grand'Henry J. (1988). La version arabe du Discours 24 de Grégoire de Nazianze. Édition critique, commentaires et traduction. In Coulie B. (Ed.), *Versiones Orientales, repertorium Ibericum et studia ad editiones curandas* (*Corpus Christianorum. Series Graeca,* 20. *Corpus Nazianzenum,* 1). Brepols : 197-291.

Grand'Henry J. (1996). *Sancti Gregorii Nazianzeni opera. Versio Arabica antiqua, I. Oratio XXI* (*Corpus Christianorum. Series Graeca,* 34. *Corpus Nazianzenum,* 4). Brepols.

Mossay J. et Cetedoc (1990). *Thesaurus Sancti Gregorii Nazianzeni. Vol. 1. Orationes, Epistulae, Testamentum* (*Corpus Christianorum. Thesaurus Patrum Graecorum)*. Brepols.

Ouersighni R. (2001). A major offshoot of the DIINAR-MBC project: *AraParse*, a morpho-syntactic analyzer for unvowelled Arabic texts. In *Proceedings of ACL/EACL*.

Tuerlinckx L. (2000). *Sancti Gregorii Nazianzeni opera. Versio Arabica antiqua, II. Orationes I, XLV, XLIV* (*Corpus Christianorum. Series Graeca,* 43. *Corpus Nazianzenum,* 10). Brepols.