Quel corpus peut aider à fonder la grammaire d'une langue pluriglossique ? Exemple de l'arabe contemporain

Catherine Pinon

Quel corpus peut aider à fonder la grammaire d'une langue pluriglossique ? Exemple de l'arabe contemporain.

Catherine PINON, IREMAM

Introduction

Cette réflexion autour de l'impact du caractère pluriglossique d'une langue sur la conception d'un corpus aurait tout aussi bien pu s'intituler : quel corpus pour fonder la grammaire d'une langue qui n'est la langue maternelle de personne ? Ou encore : quel corpus de référence pour une langue englobant plusieurs états historiques et de nombreuses variétés, utilisée sur un vaste territoire ? Comment intégrer la pluriglossie dans les études linguistiques de corpus ? Ces questions soulèvent le problème auquel tout linguiste arabisant est confronté lorsqu'il veut étudier la langue arabe, objet vivant, pluriel et mouvant qu'il est malaisé de définir.

La pluriglossie devient un problème d'ordre méthodologique lorsque l'on s'attèle à constituer un corpus représentatif d'une langue dotée d'une telle caractéristique. Habituellement, la population que le corpus vise à représenter est clairement définie et plus elle est restreinte, moins elle pose de problèmes. Ainsi, un corpus de l'arabe du Coran, de l'arabe de presse ou encore de l'arabe de la littérature semble *a priori* assez simple à définir (mais pas forcément à constituer). Il suffit, pour les deux derniers exemples, de définir la période et l'origine géographique des textes. Mais, lorsque l'on tend à représenter *l'arabe*, une multitude de questions nous assaillent, que nous pourrions résumer en une seule : qu'entend-on désigner par le terme *langue arabe* ?

Comme toutes les études adoptant la méthodologie du corpus, il convient avant tout de définir précisément la langue objet de cette recherche, mais aussi le cadre théorique dans lequel nous nous situons et les objectifs qu'elle se donne. La perspective adoptée ici est tout à fait didactique : le corpus constitué a pour but de servir de référence pour élaborer des outils adéquats et pertinents pour l'enseignement de l'arabe, langue vivante étrangère. Cette finalité nous amène à définir la population visée par l'élaboration du corpus et la forme de celui-ci d'une certaine manière, en faveur de laquelle nous argumenterons ici.

Spécificités de la langue arabe

Lorsque l'on désigne la langue arabe, soit on le fait de manière large en parlant de "l'arabe", sans plus de distinction, soit on accole un adjectif censé préciser de quel arabe on parle. Ainsi, entend-on parler de l'arabe *coranique*, *classique, littéraire, moderne, de presse, moyen, dialectal, littéral, standard, etc.* Parfois même, ces étiquettes qui désignent tantôt un état (un stade historique de la langue), tantôt une variété (une forme de la langue), sont combinées : arabe standard moderne (*modern standard arabic*), arabe littéraire classique ou moderne, voire "arabe classique moderne"[1].

L'arabe est un cas d'école en matière de pluriglossie. Ce phénomène est l'un des facteurs faisant de cette langue une réalité difficilement appréhendable, un objet d'étude qui soulève de nombreuses questions d'ordre méthodologique et dont l'analyse s'avère souvent délicate pour le linguiste. Nous parlons ici de l'arabe moderne, en tenant pour acquis le fait que cette langue, comme toutes les autres, évolue et connaît donc actuellement un état moderne : c'est là une évidence pour le linguiste, mais ne l'est pas forcément pour le commun des locuteurs arabophones.

L'arabe est doublement pluriglossique. On peut déceler une *pluriglossie diachronique* dans la langue actuelle, car des états antérieurs y survivent en des proportions assez conséquentes pour faire partie intégrante de la langue "moderne". C'est le cas principalement de l'arabe pré-classique et classique représentés par le Coran et les textes religieux, ou encore par la poésie et les œuvres littéraires qui tiennent une grande place dans la culture arabe actuelle. Depuis son plus jeune âge, un Arabe entend ou apprend par cœur d'importants passages du Coran, des *ḥadīṯ* (dires du prophète servant d'*exemplum* pour les musulmans), des prières ou encore des poèmes archaïques. Ainsi, cet état historique de l'arabe est-il toujours bien présent ; le passage par l'école le place d'ailleurs au-dessus de tout autre état ou variété, car, à grand renfort de grammaire, l'on enseigne sa "perfection".

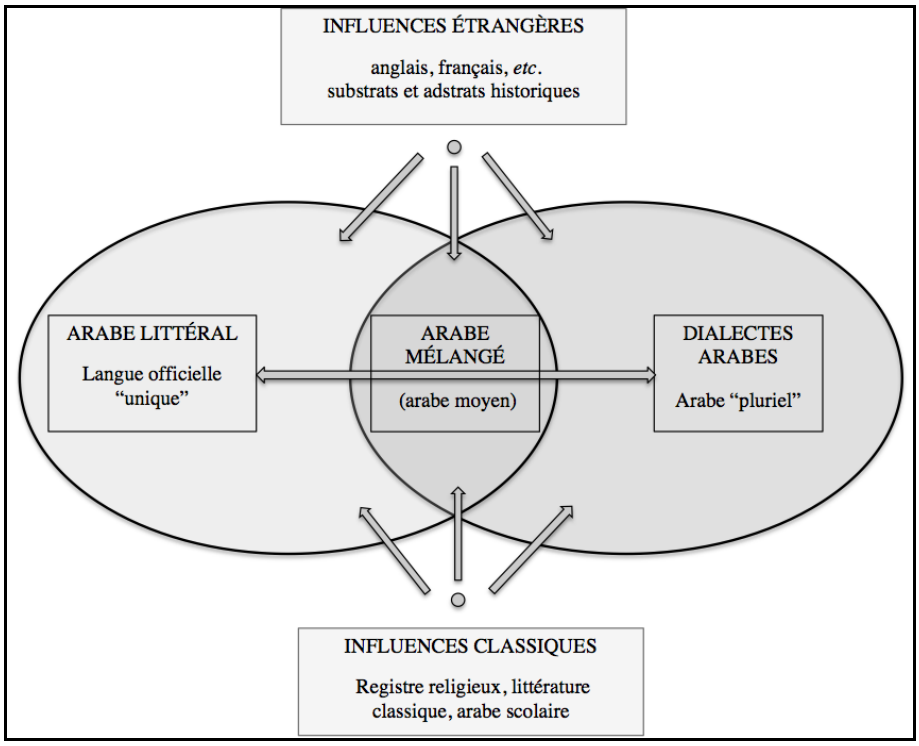
D'autre part, la *pluriglossie* que l'on pourrait qualifier de *synchronique* est bien connue : la langue maternelle de tout Arabe est son dialecte. Les dialectes arabes sont habituellement regroupés en familles : arabe maghrébin, nilotique, syro-libanais, péninsulaire ; d'autres dialectes, comme le *ḥasaniyya* parlé en Mauritanie, sortent de cette typologie. Leur répartition n'est pas liée aux frontières politiques actuelles. De ce fait, l'étendue du domaine arabe sur un vaste territoire où se parlent de nombreux dialectes ainsi que d'autres langues (le berbère, le kurde, l'hébreu, *etc.*) contribue à enrichir le nombre de ces variétés. Aux arabes dialectaux s'oppose, selon le point de vue, l'arabe *classique* ou un arabe *littéral*. Le premier point de vue est traditionnel, ou plutôt devrait-on dire idéologique[2] : à l'origine, il n'y avait qu'un arabe, l'arabe *classique*, dont le Coran représenterait la perfection inimitable. Ensuite, cet arabe, du fait des contacts avec les non-arabophones mais aussi à cause de l'imperfection humaine, s'est "dégradé" et a donné naissance aux multiples dialectes, selon la théorie de la « corruption de la langue » (*fasād al-luġa*) chère à la tradition arabe. Cette vision mythologique de l'arabe ne tient pas à l'épreuve de la linguistique historique, mais prédomine encore dans le monde arabe, où l'arabe classique est survalorisé et écrase les dialectes à tel point que beaucoup de locuteurs pensent que leur dialecte « n'est pas une vraie langue », « n'a pas de grammaire », *etc.* Il s'en suit une attitude face à la langue que d'aucuns ont qualifié de *schizophrénie de l'arabe* ou encore de *schizoglossie*[3]. Notons cependant que selon les pays, les locuteurs considèrent plus ou moins mal leur dialecte. Ainsi, au Maroc et en Egypte, l'emploi du dialecte (y compris à l'écrit) se fait-il plus naturellement que dans les autres pays.

Quel arabe enseigner ?

Cette tension entre arabe dialectal et arabe *littéral* est aussi présente à l'extérieur du monde arabe : les professeurs d'arabe langue étrangère y sont confrontés et doivent adopter une position claire. En dehors d'un pays arabe, il est communément admis d'enseigner l'arabe *littéral*, langue officielle de 22 pays représentant environ 300 millions de locuteurs. Ce terme de *littéral* présente plusieurs avantages, en premier lieu celui de ne pas référer à l'état classique de la langue. Il signifie aussi que ce type d'arabe est celui de la chose écrite, insistant sur son usage presqu'exclusivement réservé à l'écrit. Principalement utilisé sous forme écrite, l'arabe littéral est en effet réservé à des contextes de production bien spécifiques (littérature, presse, écrits administratifs ou officiels comme les manuels scolaires, les notices d'utilisation). Son oralisation, rarement spontanée, est souvent monologique (journal télévisé ou radiophonique, publicité). Cette langue, rappelons-le, n'est la langue maternelle de personne.

Il peut paraître paradoxal d'enseigner une langue vivante qui n'est la langue maternelle de personne. Quoi qu'il en soit, il est délicat d'enseigner l'arabe dans sa multiplicité d'états et de variétés. C'est en partie ce qui explique que les typologies de cette langue soient nombreuses. Bien souvent, le fait d'élaborer une typologie de la langue procure un sentiment de sécurité : on se réfère à des catégories théoriquement bien déterminées plutôt que de se confronter au mélange d'état et de variétés présent en arabe moderne. Certains professeurs vont donc décider d'enseigner tel type d'arabe, exclusivement : le *standard*, sans parler des dialectes ; le *classique*, sans aborder les évolutions. Il nous semble que l'établissement de catégories regroupant les différentes variétés d'arabe est une attitude qui peut apparaître réductrice en cela qu'elle ne prend pas en compte le fait que l'arabe est avant tout une langue mélangée.

Pour représenter la vision que nous avons de la langue arabe, nous proposons le schéma suivant[4], où l'on peut constater que les différentes variétés interagissent entre elles. Ainsi, l'arabe que nous avons nommé littéral, à gauche, connaît de multiples influences (langues étrangères, langues du domaine arabe, mais aussi états anciens de la langue, dialectes, *etc.*) : l'enseigner oblige, pour respecter un principe de réalisme linguistique[5], à reconnaître ces composantes comme inhérentes à la langue elle-même.



Représentation schématique de l'arabe contemporain

Une lacune disciplinaire entraînant un besoin didactique

Traditionnellement, les grammaires de l'arabe s'élaborent par accrétion. Ce phénomène basé sur la reprise des travaux antérieurs, avec suppressions et ajouts, a aussi été reproduit par les grammairiens orientalistes[6]. Ainsi, il est encore peu habituel, pour élaborer une grammaire de l'arabe, de se confronter aux textes, même si plusieurs grammaires d'arabe moderne fondées sur corpus ont vu le jour[7]. Actuellement, pour les linguistes arabisants, cette lacune n'est pas due à une position idéologique de refus des évolutions de l'arabe. Il s'agit plutôt d'une lacune disciplinaire héritée de la tradition grammaticale, mais qui s'estompe avec l'essor de la linguistique de corpus. Dorénavant, le frein à la description de la langue contemporaine est davantage causé par des raisons pratiques : le recueil et le traitement des données est malaisé (nous y reviendrons). Mais, au-delà de la simple description de la langue, c'est une analyse minutieuse de son fonctionnement qui est nécessaire.

En effet, une fois la langue qu'il enseigne bien définie, le professeur d'arabe se heurte à un problème d'ordre pratique. S'il veut faire en sorte que son enseignement de la grammaire soit au plus proche des emplois réels, il ne trouvera pratiquement pas de grammaires de référence pour l'aider à construire son cours. Le manque de grammaires de référence empêche certains professeurs de moderniser le contenu de leur cours de langue arabe (grammaire, expression écrite, *etc.*), car tel fait de langue qui s'écarterait de la norme classique ne figurant pas dans une grammaire, il n'a pas acquis un statut assez officiel lui permettant d'être enseigné. La grammaire – le livre – joue un rôle considérable dans les études arabes : elle sert souvent d'argument décisif et incontestable pour rejeter un usage pourtant attesté en nombre. En grammaire arabe, ce recours à l'argument d'autorité du linguiste est incontournable. Il est donc nécessaire, pour que l'enseignement de la langue puisse évoluer, d'élaborer des outils adéquats à la langue enseignée.

La linguistique de corpus : une réponse méthodologique

La linguistique de corpus apparaît comme une réponse appropriée aux questionnements liés à l'élaboration d'une grammaire réaliste de la langue telle qu'elle est actuellement employée. La question primordiale est donc de définir le corpus qui devrait servir de base à la constitution d'une telle grammaire. Les ouvrages existants les plus complets ont été réalisés à partir de corpus, mais la composition de ceux-ci pourrait leur être reprochée, car ils n'extraient bien souvent leurs données que de quelques sources ressortissant au même genre (quelques romans, quelques journaux, *etc.*). La

grammaire d'arabe littéral moderne la plus sérieuse, à notre sens, est celle de E. Badawi, M. Carter et A. Gully (2004). Elle est fondée sur corpus et présente l'aboutissement d'un projet ambitieux, voulant réunir théoriquement tous les types d'écrits modernes, du graffiti à la littérature. Mais, concrètement, la majeure partie des sources utilisées provient de la presse. Par ailleurs, si nous devons saluer l'apparition de ce type d'ouvrages parce qu'elle souligne un changement de paradigme dans les études portant sur la grammaire arabe, nous pouvons tout de même relever certains défauts. Le principal est l'emprise du classicisme, toujours latent, qui empêche une recatégorisation des faits linguistiques lorsque celle-ci est nécessaire. Une autre critique pourrait être émise lorsque les exemples ne sont pas référencés systématiquement. Ces défauts de forme disparaîtront sans doute à mesure que ce genre de travail sera développé.
Quoi qu'il en soit, si la linguistique de corpus constitue une réponse méthodologique au problème ci-dessus exposé de l'inadéquation descriptive de la majorité des grammaires d'arabe moderne[8], il n'en reste pas moins que cette méthode doit être affinée et constamment améliorée.

Un corpus de référence pour l'arabe : une réelle nécessité ?

Si les corpus réunissant de la matière textuelle en langue arabe se multiplient, il n'existe pas, à ce jour, de corpus de référence pour cette langue. Nombreux sont les chercheurs qui font ce constat et justifient ainsi la constitution de leur propre corpus.
L'engouement actuel pour le corpus, devenu le « sésame obligatoire » à toute étude, pour reprendre l'expression de D. Mayaffre, précipite bien souvent le chercheur vers la conception de cet objet, sans qu'il ne se pose une question pourtant fondamentale : ai-je réellement besoin d'un corpus, et si oui, de quel corpus ? On passe souvent directement aux questions formelles : quelle taille doit-il faire ? Quels types de texte doit-il regrouper ? L'idée d'utiliser des corpus déjà existants ne se pose même pas, tant on est accaparé par la construction de cet objet incontournable, et conforté dans cette attitude par le fait que chaque corpus ne peut être élaboré qu'en fonction des objectifs de recherche. Les critères sont-ils chaque fois si différents qu'un corpus ne puisse être réutilisé ?
Si, dans le cadre d'une grammaire réaliste de l'arabe contemporain, l'élaboration d'un corpus à la hauteur des ambitions du projet semble incontournable, nous nous devons d'émettre une réserve. Nous constatons que les chercheurs les plus productifs ont élaboré de "petits" corpus leur permettant de les analyser et d'en tirer des résultats[9], alors que les chercheurs qui ont œuvré à l'élaboration d'un vaste corpus en sont souvent resté au stade de la théorisation du corpus ou, du moins, peut-on estimer que les résultats obtenus n'ont pas été à la hauteur des efforts prodigués pour le corpus en lui-même[10]. Ainsi, ce dernier finit-il par apparaître comme une fin en soi, alors qu'il ne devrait être qu'un moyen, un objet heuristique, un adjuvant au travail du linguiste. A l'inverse, certains chercheurs ont réutilisé les corpus déjà constitués pour créer des bases de données diversifiées[11]. Par ailleurs, les différents domaines de la linguistique ne sont pas également représentés : si les recherches techniques portant sur le traitement automatique de la langue ou les études ayant pour objet le lexique sont nombreuses, celles traitant de la syntaxe et de la sémantique sont les moins courantes.
L'autre avantage qu'il y a à faire l'état de la recherche avant de se lancer dans l'élaboration d'un corpus est double : non seulement cela permet d'accéder à des ressources textuelles, mais surtout de bénéficier de l'expérience de ses prédécesseurs qui ne manquent pas, lorsqu'ils présentent leurs corpus, de rappeler quelques problèmes techniques et pratiques rencontrés.
Que l'on choisisse de réutiliser des données déjà collectées ou que l'on décide de partir de zéro, il est évident qu'un tel projet doit être mené par une équipe de chercheurs, regroupant linguistes, statisticiens et informaticiens.

Le corpus idéal

En vue d'établir une grammaire d'arabe littéral contemporain la plus complète et la plus réaliste possible, il faudrait pouvoir se baser sur un corpus de référence. Quelles devraient être, dans l'idéal, les propriétés d'un tel corpus ?

Définition de la population

Nous avons indiqué plus haut les raisons qui exigent que la définition de la "population", au sens statistique du terme, à savoir dans ce cas précis l'arabe littéral contemporain, soit mûrement réfléchie.

Les notions de niveau de langue, de variétés et de genres devront notamment être discutées afin de sélectionner les sources en vue de constituer un corpus représentatif, cohérent et pertinent. Quels facteurs faut-il prendre en compte ?

- Facteur temporel : les limites peuvent être définies arbitrairement, *a priori* (les données doivent être produites à partir de telle année) ou ajustées au cours de la recherche des données disponibles en fonction de leur accessibilité (élargissement ou restriction de la période).
- Facteur diatopique : une représentativité exhaustive amènerait à collecter des données provenant de tous les pays arabes, mais aussi plus largement à accepter les données produites à l'extérieur du monde arabe (pays musulmans, terres d'immigration, pays d'accueil de délocalisations, *etc.*).
- Facteur générique : un inventaire des différentes productions en arabe littéral devrait être fait pour permettre d'établir une hiérarchie des genres ou types de textes[12] devant figurer dans le corpus. Outre la littérature et la presse, genres majeurs qu'il conviendrait de subdiviser, on ne doit pas oublier les productions plus techniques (manuels, notices, articles, notes de service, *etc.*) ni celles ressortissant à un nouveau mode de communication (emails, billets postés sur les blogs, forums, clavardage, *etc.*).
- Facteurs relatifs au mode et au support de production : bien évidemment, il faudrait que figurent des données écrites et des données orales. Mais au-delà de la question de l'écrit et de l'oral, c'est aussi le support de production qui doit être précisé. Il faudrait par exemple veiller à ce que les données ne soient pas toutes initialement numériques (ce qui implique un coût de traitement des données papier).

Taille et typologie des données textuelles

Les questions de la taille que doit faire le corpus et des différents genres et types de textes qui doivent y être intégrés découlent directement du souci de représentativité. Deux écoles s'opposent théoriquement : pour la première, on amasse une quantité de données considérables, un peu au hasard, en arguant du fait que la quantité de mots en elle-même assurera une certaine représentativité, mais alors, la question du traitement et de l'analyse de ces données devient cruciale ; pour la seconde, on choisit de sélectionner peu de données au regard de critères nombreux et précis. Le problème qui survient alors est que, plus l'on multiplie le nombre de ces critères, plus la taille du corpus augmente en conséquence.

Il nous semble important de prendre en compte deux choses : veiller, d'une part, à l'équilibre entre corpus et outils, pour que le corpus soit explorable[13] et analysable de manière satisfaisante, quelle que soit sa taille, en fonction des outils à disposition des chercheurs. Limiter, d'autre part, son avidité à la collecte des données (surtout lorsque celles-ci sont facilement accessibles), en n'oubliant pas que l'accroissement des faits de langue n'est pas proportionnel à leur nombre. A ce propos, M. Van Mol[14] estime qu'à partir de 200.000 mots, l'accroissement du vocabulaire est quasi-nul. Il n'est donc pas nécessaire de collecter des données à l'infini, le rapport entre la perte de temps occasionnée par la collecte des données, leur traitement et leur analyse, et le gain de vocabulaire nouveau ou de structures originales n'étant pas fructueux. Bien évidemment, les chercheurs prendront en compte ce facteur taille de manière différente selon qu'ils travaillent sur un lexique spécialisé, sur une particule très courante, sur la phonologie, la morphologie ou la syntaxe de la langue.

Si, finalement, la taille n'apparaît pas comme le moyen fondamental d'atteindre à la meilleure représentativité qui soit, la notion de genre nous paraît beaucoup plus importante. Nombreux sont les chercheurs qui ont démontré l'importance de fonder le corpus sur une typologie des genres (notion comprise dans une acception plus large que ce qu'elle recouvre dans la classification traditionnelle des textes)[15]. C'est en diversifiant au maximum les genres que l'on peut obtenir des faits de langue plus nombreux ; ce n'est d'ailleurs pas seulement vrai du lexique, mais aussi de la syntaxe. Pour déterminer les différents genres qui doivent être représentés, on peut s'inspirer des typologies détaillées existantes, comme celles de S. Atkins et *alii.* ou de J. Sinclair, mais aussi utiliser les classifications déduites *a posteriori*, comme celles proposées par D. Biber. Les deux typologies sont complémentaires et visent à affiner la sélection et le regroupement des données textuelles.

Vers un corpus réalisable : la gestion des problèmes pratiques

Si la définition du corpus représente une étape de réflexion aussi minutieuse qu'indispensable, elle ne résout en rien les problèmes pratiques auxquels tout linguiste élaborant son corpus sera confronté. L'accès aux données, leur collecte, leur traitement, leur homogénéisation, mais aussi leur analyse vont

parfois poser problème, au point de susciter de nouvelles réflexions, voire de revenir sur certains choix préétablis théoriquement. En fait, le réajustement constant de la forme du corpus est l'une des caractéristiques de cet objet d'étude "donné-construit". C'est le processus cyclique décrit par D. Biber et repris par d'autres linguistes : conception du corpus > analyse des données et résultats > bilan critique > amélioration du corpus.
Les points abordés ci-dessous ne sont certainement pas les seuls à remettre en question la conception et l'élaboration des corpus en linguistique arabe : nous en évoquons certains et en omettons d'autres, en particulier les problèmes informatiques et techniques, car ils dépassent notre compétence ; le but de cette énumération étant d'engager la réflexion.

Variétés d'arabe majorées ou minorées *de facto*

Du fait des contraintes matérielles (accessibilité des ressources, obtention des droits d'auteur, coût de la saisie des textes écrits non numériques ou de la transcription des discours oraux, développement de logiciels adaptés pour le traitement du corpus, *etc*.), certaines variétés ou certains emplois de la langue arabe sont sur-représentés de facto dans les études linguistiques, alors que d'autres sont sous-représentés. D'une manière générale, l'arabe écrit est privilégié par rapport à l'oral, de même que l'arabe classique ; pour les textes contemporains, l'arabe de presse et, dans une moindre mesure, celui de la littérature sont les plus étudiés, avec une prédisposition pour les sources égyptiennes et syro-libanaises. Les dialectes sont assez peu étudiés à grande échelle. D'autres variétés nous semblent sous-représentées dans les études linguistiques, notamment l'arabe technique, scientifique ou administratif, l'arabe "privé" des correspondances, blogs, forums, sites internet, émissions télévisuelles, *etc*. ; l'arabe "mélangé" (standard / dialecte) et plus généralement l'arabe non-standard, mais aussi l'arabe parlé au-delà des frontières du monde arabe.
Un corpus de référence tendrait à rétablir un certain équilibre dans la représentation de ces variétés, mais dans la pratique les contraintes seraient telles que le travail apparaîtrait comme titanesque et demanderait que les fonds alloués au projet soient considérables.

Le règlement des questions juridiques

L'obtention des droits d'auteurs s'avère être une étape chronovore et déconcertante pour les chercheurs, qui la considèrent bien souvent comme une perte de temps. Si, dans le cadre de recherches ponctuelles et particulières (comme un travail de thèse), à la diffusion très restreinte, nous estimons que l'on peut se permettre de passer outre les droits d'auteurs pour "raisons scientifiques", créer un corpus de référence accessible aux chercheurs et permettant d'élaborer différents outils ou de mener à bien des recherches théoriques implique nécessairement la résolution de ce problème.
Il convient donc, au cours de l'élaboration du corpus, d'identifier les différents types de données collectées et de se renseigner sur le régime juridique de chacun d'eux, qui dépend en général des législations nationales. Dans de nombreux cas, des autorisations seront nécessaires et devront être négociées en fonction notamment de l'utilisation prévue du corpus. Il ne faut pas oublier non plus que le corpus lui-même, une fois constitué, ainsi que les travaux qui en seront issus, devront être protégés.

La collecte des données

Les données à intégrer sont de trois ordres : soit disponibles sous un format numérique, soit écrites mais non numériques (manuscrits ou tapuscrits), soit orales. Pour chaque type de données, le travail de récupération, de nettoyage et d'homogénéisation ne sera pas le même.
Les sources numériques peuvent apparaître comme les plus simples à traiter. Néanmoins, elles poseront tout de même de nombreux problèmes, en premier lieu celui de l'encodage. Les ressources numériques (littérature, presse, articles ou toute autre sorte de texte publiés sur le Web, sites Internet, blogs, emails, fichiers PDF extractables, *etc.*), selon qu'elles revêtent un caractère très officiel ou privé, qu'elles aient fait l'objet d'un travail minutieux ou dans l'urgence, vont comporter plus ou moins d'abréviations (quoi que cet aspect soit moins développé en arabe que pour d'autres langues), de coquilles ou encore d'habitudes "fautives" systématisées[16]. Quelle attitude adopter ? Doit-on relire tous les textes en corrigeant ces erreurs, ou les conserver en estimant que la quantité de données palliera ces défauts qui seront considérés comme statistiquement négligeables ? Tout dépendra bien évidemment de l'usage qui sera fait du corpus. Par ailleurs, d'autres problèmes peuvent se poser, comme la redondance d'articles de presse, disponibles dans plusieurs journaux en ligne, ou encore

plus trivialement les questions de connexion (parfois interrompue durablement), de mise à jour ou encore de censure (toujours très opérante). On peut aussi s'interroger sur le statut à conférer aux données en langue arabe mais transcrites en caractères latins.
Les sources écrites non numériques ne posent pas les mêmes problèmes de collecte selon qu'il s'agit de manuscrits ou de tapuscrits. Pour les manuscrits, si des logiciels de reconnaissance graphiques en arabe ont été testés, aucun, à notre connaissance, n'est assez performant pour permettre la numérisation de documents écrits à la main. La solution la plus simple consiste à saisir le texte, ce qui a été en grande partie réalisé pour les œuvres classiques et qui commence à apparaître pour les textes littéraires modernes[17]. Ceci s'avère en réalité plus coûteux que long. Pour les tapuscrits, il faut recourir au système de reconnaissance optique des caractères (OCR), ce qui s'avère long et coûteux car ces logiciels ne sont pas parfaitement performants pour l'arabe, d'autant plus que souvent, la qualité d'impression du document original est médiocre. Il vaut mieux saisir de nouveau ces textes. Pour la littérature, contacter les auteurs arabes s'avère généralement inutile, car ils sont encore nombreux à composer leurs œuvres sur papier, la maison d'édition se chargeant de la saisie. De fait, beaucoup ne possèdent pas eux-mêmes de versions électroniques de leurs travaux, et les maisons d'éditions sont plutôt réticentes à les fournir. Bien évidemment, ce n'est pas le cas de la nouvelle génération d'écrivains qui mettent leurs textes à disposition sur Internet, ce qui n'est pas sans poser de problèmes juridiques en cas de collecte lorsque ceux-ci ont par ailleurs été édités sur papier.
Nous ne nous étendrons pas sur les sources orales, car les problèmes relatifs à leur transcription sont bien connus. Sans même parler de leur obtention, transcrire des enregistrements oraux est si long et coûteux que la plupart des corpus excluent totalement ce type de données. Les seuls à recourir abondamment aux corpus oraux sont les sociolinguistes, mais ils travaillent alors sur un parler très précis (le parler des jeunes dans tel quartier, le dialecte de tel village, *etc.*).

"Nettoyage", homogénéisation et encodage des données

Cette étape, fondamentale car déterminant la qualité de traitement des données par la machine, est assez complexe. Selon M. Van Mol (2007 : 300), il est urgent de résoudre les problèmes de codage des données, pour parvenir à adapter totalement les logiciels et utilitaires à toutes les nécessités imposées par l'arabe. Les pertes sont encore trop nombreuses, que ce soit lors du passage des données d'un ordinateur de type Mac à un PC, ou lors de leur enregistrement en texte brut. Parfois, ces problèmes d'encodage, s'ils ne sont pas visibles sur l'écran, se répercutent à l'impression. Quoi qu'il en soit, il semble pertinent d'élaborer un protocole précis présidant au captage de la matière textuelle, non seulement pour gagner du temps, mais surtout pour s'assurer que les données recueillies pourront effectivement être traitées par les logiciels choisis.

Les particularités graphiques et morphosyntaxiques de l'arabe

Sans entrer dans des considérations techniques trop détaillées, nous désirons ici évoquer deux problèmes liés aux particularités graphiques et morphosyntaxiques de l'arabe.
Tout d'abord, rappelons qu'en arabe, seules les voyelles longues sont notées. Les voyelles brèves ou encore la gémination ne sont en général pas notées, mais peuvent l'être au moyen de signes que l'on rajoute au-dessus ou en-dessous des mots. C'est ce que l'on nomme la vocalisation d'un texte. En général, lorsque l'on accède à des données, elles ne sont pas vocalisées, ou alors partiellement et parfois de manière fautive. Bien souvent, si la vocalisation passe à l'œil nu, il apparaît qu'elle est mal faite, numériquement (la voyelle est codée avant la consonne par exemple)[18]. Cette question de vocalisation est cruciale dans la mesure où elle permet de lever de très nombreuses ambiguïtés graphiques (en arabe, dans un texte non-vocalisé, le taux de mots pouvant être lus et analysés de différentes manières est assez élevé).
Le second problème a trait au caractère concaténatoire de la graphie arabe : un mot graphique peut être composé d'une préposition affixée, du nom ou d'un verbe et d'un pronom suffixé. Dans le tableau ci-dessous, on présente successivement des mots non-vocalisés, vocalisés puis décomposés.

بكرته	ليكتبها
بِكُرَتِه	لِيَكْتُبَها
بـ/ كرة/ ه	لـ/ يكتب/ ها
avec sa balle	*pour qu'il l'écrive*

Exemple de mots graphiques arabes

La morphologie de l'arabe va se révéler plus ou moins problématique selon le niveau d'analyse visé par le chercheur. Plusieurs outils ont été mis au point pour l'analyse automatique du mot arabe, selon deux méthodes différentes, analysant le mot graphique arabe soit en termes de racine et de schèmes, soit en termes de préfixe, noyau et suffixe, en recourant à la lemmatisation[19].

Les outils disponibles

Il existe à l'heure actuelle trop peu d'outils performants mis à la disposition d'un large public sur Internet. Les concordanciers sont relativement nombreux, mais il est plus difficile de se procurer un étiqueteur morpho-syntaxique par exemple, pourtant indispensable pour mener à bien des études approfondies[20]. Les équipes de chercheurs où linguistes et informaticiens travaillent de concert mettent en général au point des outils performants adaptés aux objectifs de leur recherche. Quoi qu'il en soit, il est incontournable de se former aux outils adéquats disponibles actuellement, car les études linguistiques s'appuyant sur des corpus "bricolés" de manière plutôt artisanale sont encore légion. Dans le cadre de l'analyse d'un corpus de référence, il est indispensable de s'aider au minimum de vocaliseur-désambiguïseur de texte automatiques et d'étiqueteurs morpho-syntaxiques, mais aussi d'avoir à disposition un logiciel spécialisé contenant différentes fonctionnalités (concordancier, recherche de segments répétés, partitions du corpus, analyses statistiques, calculs et représentations statistiques multidimentionnels, *etc.*).

Conclusion

Une fois de plus, un article aura été consacré aux questions théoriques relatives à la constitution d'un corpus. Revenons tout de même sur le projet fondateur, au sujet duquel nous n'avons fait ici qu'évoquer quelques difficultés de réalisation car ce projet ambitieux a pour principal obstacle la technique (des outils trop épars et pas assez performants). Un corpus représentatif d'arabe contemporain, une fois constitué, figurerait un instantané de la langue arabe actuelle. Cette photographie que les linguistes s'attacheraient à rendre la plus nette possible, pourrait par la suite servir de comparaison dans le cadre d'études de linguistique historique. Le développement d'outils performants pourrait amener à faire de ce corpus de référence un corpus ouvert, sans cesse augmenté de nouvelles données : une sorte de base de données de veille de la langue qui compenserait le caractère fugace du travail sur la langue "contemporaine". Utilisée pour élaborer des outils didactiques (grammaires, dictionnaires, cahiers d'exercice, lexiques, *etc.*), une telle base de données amènerait à redéfinir le concept de norme d'une langue. On en viendrait peut-être à élaborer des grammaires plus réalistes, à admettre une norme évolutive et peut-être que ceci contribuerait enfin à désacraliser, démystifier et désidéologiser la langue arabe. Le rôle du linguiste, du didacticien et de l'enseignant d'arabe prendrait alors tout son sens.

Bibliographie

Al-Ansary, S., Nagi, M. et Adly, N. (2008), « Building an International Corpus of Arabic (ICA) : Progress of Compilation Stage ». Récupéré le 29 novembre 2009 de http://www.bibalex.org/isis/UploadedFiles/Publications/Building%20an%20Intl%20corpus%20of%20arabic.pdf

Al-Ansary, S., Nagi, M. et Adly, N. (2008), « Towards Analyzing the International Corpus of Arabic (ICA) : Progress of Morphological Stage », Bibliotheca Alexandrina. Récupéré le 29 novembre 2009 http://www.bibalex.org/isis/uploadedfiles/publications/morphological_analysis_of_ica_finalx.pdf

Al-Sulaiti, L. et Atwell, E. (2003), « The Design of a Corpus of Contemporary Arabic (CCA) », *School of Computing research report 2003.11*. Récupéré le 29 novembre 2009 http://www.comp.leeds.ac.uk/research/pubs/reports/2003/2003_11.pdf

Al-Sulaiti, L. et Atwell, E. (2005), « Extending the Corpus of Contemporary Arabic », *Proceedings of Corpus Linguistics*. Récupéré le 29 novembre 2009 de http://www.comp.leeds.ac.uk/eric/alsulaiti05cl.pdf

Atkins, S, Clear, J. et Ostler, N. (1992), « Corpus Design Criteria », *Literary and Linguistic Computing*, vol. 7, n°1, pp. 1-16

Badawi, E. Carter, M. et Gully, A. (2004), *Modern Written Arabic*. *A comprehensive Grammar*. London and New York, Routledge.

Biber, D. (1993), « Representativeness in Corpus Design », *Literary and linguistic Computing*, vol. 8, n°4, pp. 243-257

Ditters, E. (1990), « Arabic Corpus Linguistics in Past and Present », *in* Versteegh, K. et Carter, M. (éds.), *Studies in the History of Arabic Grammar II*, pp. 129-141

Habert, B. (2000), « Des corpus représentatifs : de quoi, pour quoi, comment ? », *in* Bilger, M. (coord.), *Linguistique sur corpus. Etudes et réflexions*. Cahiers de l'Université de Perpignan n° 31, Presses Universitaires de Perpignan, pp. 12-58

Habert, B. (2004), « Portrait de linguiste(s) à l'instrument ». Récupéré le 24 novembre 2009 de http://www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html

Habert, B., Fabre, C. et Isaac, F. (1998), *De l'écrit au numérique. Constituer, normaliser et exploiter les corpus électroniques.* Paris, InterEditions, Masson

Habert, B., Nazarenko, A. et Salem, A. (1997), *Les linguistiques de corpus*. Paris, Armand Colin

Malrieu, D. (2005), « Domaines, champs génériques, temps et personnes », *in* Williams, G. (dir.), *La Linguistique de corpus*, Actes des deuxièmes journées de la linguistique de corpus, Lorient 12-14 septembre 2002, Presses Universitaires de Rennes, pp. 115-129

Malrieu, D. et Rastier, F. (2001), « Genres et variations morphosyntaxiques », *Traitement Automatique des langues*, vol. 42, n°2, p. 548-577. Récupéré le 25 novembre 2009 de http://www.revue-texto.net/Inedits/Malrieu_Rastier/Malrieu-Rastier_Genres.html

Mayaffre, D. (2005), « Rôle et place des corpus en linguistique : réflexions introductives », *Texto !*, vol. X, n°4. Récupéré le 26 novembre 2009 de http://www.revue-texto.net/Reperes/Themes/Mayaffre_Corpus.html

Pinon, C. (2011), « La grammaire arabe : entre théories linguistiques et applications didactiques », *Synergies Monde arabe* n°7, pp. 75-86. Disponible en ligne : http://ressources-cla.univ-fcomte.fr/gerflint/Mondearabe7/mondearabe7.html

Sinclair, J. (1996), « Preliminary Recommandations on Text Typology », *EAGLES* EAG-TCWG-TTYP/P. Récupéré le 9 octobre 2010 de http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html

Van Mol, M. (2003), *Variation in Modern Standard Arabic in Radio News Broadcasts. A synchronic Descriptive Investigation in the use of complementary Particles*. *Orientalia Lovaniensia Analecta* 117, Leuven, Peeters

Van Mol, M. (2007), «Arabic and the Computer : Possibilities and perspectives for scientific research and educational purposes », *in* Hamdani, A., Lachhab, K. et Erradi, M. (éds.), *Traitement automatique de la langue arabe / Arabic Language Processing*, Actes du colloque Proceedings, juin 2006, Université Mohammed V – Souissi, Institut d'Etudes et de Recherches pour l'Arabisation, pp. 299-317

---

[1] Dans cette expression, *classique* désigne la variété standard de la langue (l'arabe tel qu'il est enseigné à l'école) et *moderne* son état. Bien souvent encore, on enseigne l'arabe *classique classique*, ou plutôt, une *sui-langue arabe classique*. Sur ces deux concepts, voir Larcher, P. et Girod, A. (1990), « Passif grammatical, passif périphrastique et catégorie d'auxiliaire en arabe classique moderne », *Arabica*, 37/2, pp. 137-150 pour le premier et Pinon, C. (2011) pour le second.

[2] Pour approfondir cette question, *cf.* Larcher (2010), « *Al-lugha al-fuṣḥâ* : archéologie d'un concept "idéolinguistique" », *REMMM* n° 124, pp. 263-278.

[3] Sur les concepts de schizoglossie ou de schizophrénie de l'arabe, nous renvoyons à Calvet, L.-J., (1999), *Pour une écologie des langues du monde* et à Choubachy, C. (2007), *Le sabre et la virgule. La langue du Coran est-elle à l'origine du mal arabe ?*

[4] Cette représentation schématique n'a pour seul intérêt que de cesser d'opposer les dialectes à l'arabe classique. Elle vise à montrer que le pendant des dialectes est cet arabe littéral moderne (quel que soit le nom qu'on lui donne, il ne s'agit dans tous les cas certainement pas de l'arabe "classique"). Le schéma est volontairement symétrique, mais si nous avions voulu représenter la situation de l'oral, l'espace des dialectes aurait représenté l'écrasante majorité ; si nous avions représenté la situation de l'écrit, le contraire se serait produit : l'espace du littéral aurait été considérable. Il en va de même pour les différentes influences, qui n'ont pas le même impact sur le littéral et sur le dialectal. Comme tout schéma, il s'agit d'une représentation abstraite : la langue, elle, s'épanouit dans un *continuum* qui ne peut être segmenté par le linguiste que pour des raisons pratiques.

[5] *Cf.* Imbert, F., (2010), « Enseigner la grammaire arabe à l'université : réforme et devoir de réalisme linguistique », *in* Aguilar, V., Pérez Cañada, L. M. et Santillan Grimm, P. (éds), Enseñanza y aprendizaje de la lengua Arabe, ARABELE2009, edit.um, pp. 47-62.

[6] Pour l'évolution de la grammaire arabe, puis arabisante, et ses conséquences au niveau de l'enseignement de la langue, nous renvoyons à Pinon (2011). Sur le rôle des corpus dans l'élaboration de la grammaire classique, lire Ditters, E. (1990).

[7] La première du genre est celle de Cantarino, V. (1974-1975), *Syntax of Modern Arabic Prose* mais tout comme celle plus récente de Buckley, R. (2004), *Modern Literary Arabic*, les textes ne proviennent que de la littérature ; la meilleure est sans conteste celle de Badawi, E., Carter, M. et Gully, A. (2004).

[8] Nous renvoyons les lecteurs désireux d'avoir des exemples de l'inadéquation descriptive des grammaires de l'arabe moderne à Pinon, C. (2011) et Sartori, M., (2010), « Pour une approche relationnelle de la conditionnelle en arabe littéraire moderne », *Arabica* 57, pp. 68-98.

[9] Nous pouvons citer en exemple Van Mol, M. (2003) qui a publié une monographie présentant le résultat de ses recherches sur les variations dans les émissions radiodiffusées, à partir d'un corpus de 300.000 mots.

[10] Soit que les chercheurs ne publient pas leurs résultats, soit que les projets se soient peu à peu éteints après la constitution du corpus, d'après ce que nous avons pu trouver par exemple sur le *International Corpus of Arabic* ou sur le *Corpus of Contemporary Arabic*, les articles développent les questions méthodologiques relatives à la constitution du corpus, mais ne présentent que quelques résultats, s'arrêtant bien souvent à la morphologie de l'arabe.

[11] D. Parkinson, professeur à Brigham Young University, a réuni de nombreux corpus dans une base de données nommée arabicorpus, accessible gratuitement sur Internet, qui permet d'effectuer des recherches dans un ou la totalité des corpus réunis (presse, Coran, *Mille et Une Nuits*, littérature et dialecte égyptien).

[12] Nous pouvons reprendre ici la distinction opérée par Biber, D. (1993 : 244-245) entre *genres* (ou *register*) et *types de texte*. Les premiers renvoient à des catégories de textes définies selon des paramètres situationnels, les seconds à des catégories de textes définies linguistiquement, identifiées sur la base de modèles linguistiques co-occurrents.

[13] *Cf.* les réflexions de Habert, B. (2000 ; 2005) qui note que les corpus sont tellement vastes maintenant qu'on ne les lit plus, mais qu'on les explore. La question des outils et des instruments utilisés par le linguiste devient donc cruciale.

[14] *Cf.* la partie consacrée à la méthodologie du corpus par Van Mol, M. (2003), en particulier p. 125.

[15] Nous renvoyons tout particulièrement à Malrieu, D. et Rastier, F. (2001).

[16] La graphie du *yā'* final dans certains écrits publiés en Egypte en constitue un bon exemple, car le *yā'* qui est normalement noté avec deux points diacritiques, est souvent noté sans ces points (ce qui fait de lui un *alif maqṣūra*), alors que le *alif maqṣūra*, un caractère qui a la même forme sans points diacritiques, se voit affublé des deux points. Dans certains romans ou dans certains journaux, l'inversion est même systématique : tous les *yā'* finaux sont notés comme des *alif maqṣūra*, et *vice-versa*, ce qui génère de nombreuses confusions qu'un homme saurait vite rétablir, mais qu'une machine ne pourra pas forcément traiter automatiquement.

[17] Dans le monde arabe, de nombreuses bases de données sont vendues sur CD ou maintenant davantage sur disques durs externes, comme *al-Maktaba aš-Šāmila* (« la bibliothèque complète »), qui contient 16.000 ouvrages classiques ressortissant aux écrits religieux, juridiques et littéraires, livrés avec un logiciel de recherche et d'exploration des textes. Les initiatives de ce type pour la littérature moderne ne sont pas aussi performantes, car il s'agit bien souvent de fichiers PDF mal numérisés et donc inexploitables. Dans certains projets, des personnes sont payées pour saisir les textes : c'est le cas par exemple pour l'édition d'un très long manuscrit de la *Sīrat Baybarṣ* entreprise à l'Institut Français du Proche-Orient de Damas par G. Bohas et K. Zakharia.

[18] Ceci a amené des chercheurs, comme D. Kouloughli, à entièrement vocaliser les textes en transcription avant de les repasser en graphie arabe, de manière à obtenir un texte parfaitement transcrit.

[19] Pour plus de détails sur ces questions techniques et une présentation des différentes méthodes, nous renvoyons à Al-Ansary, S., Nagi, M. et Adly, N. (2008).

[20] Parmi les outils conçus pour l'arabe accessibles en ligne, mentionnons : aConCorde (http://www.andy-roberts.net/software/aConCorde/index.html), arabicorpus (http://arabicorpus.byu.edu/index.php), qamus (http://www.qamus.org/), Penn Arabic Treebank (http://www.ircs.upenn.edu/arabic/), kawâkib (http://www.ifao.egnet.net/kawakib/), Root search engine (http://www.comp.leeds.ac.uk/cgi-bin/scmss/arabic_roots.py), AminePlatform (http://sourceforge.net/projects/amine-platform/). D'autres outils, comme Sarfiyya (http://www.ifao.egnet.net/axes/ecritures-langues/tal-arabe/automatesarabes/automatesarabes-outils/) ou encore la base de données DIINAR (http://silat.univ-lyon2.fr/Presentation%20DIINAR.html) ne font pas l'objet de diffusions.