[Ref. : Alrahabi, Motasem et Dichy, Joseph, « Levée d'ambigüité par la méthode d'exploration contextuelle: la séquence *'alif-nûn* (ان) en arabe », in Ghenima, Malek, Ouksel, Aris et Sidhom, Sahbi (eds.), *Systèmes d'Information et Intelligence Economique, 2ème Conférence Internationale (SIIE 2009),* organisée par l'université de Nancy, France et l'université de la Manouba, École supérieure de commerce électronique (ESCE), Tunis, Tunis, Hammamet, 12-14 février 2009, IHE éditions, p. 573-585.]

Levée d'ambigüité par la méthode d'exploration contextuelle: la séquence *'alif-nûn* (ان) en arabe

Motasem Alrahabi

Laboratoire LALLIC
(Langages, Logique, Informatique, Cognition et Communication), **Université de Paris-Sorbonne**

motasem.alrahabi@paris4.sorbonne.fr

Joseph Dichy

Université Lumière-Lyon 2 et
Laboratoire ICAR (Interactions, Corpus, Apprentissages, Représentations - UMR 5191 CNRS/Lyon 2 - ENS-LSH)
joseph.dichy@univ-lyon2.fr

Abstract

In the processing of Arabic, one of the main well-known problems stems from the fact that usual writing is partly or totally « un-vowelled », and also from the omission of the *hamza* sign at the beginning of word-forms. This considerably increases ambiguity due to homography as well as to lexical polysemy. In the *'alif-nûn* (ان) written sequence, for instance, the two points above entail the possibility of referring to five different tool-words, belonging to various grammatical fields. In this paper, we resort to the approach of NLP known as Contextual Exploration Processing (Desclés et al., 1991 ; Desclés 2006). It is based on the linguistic analysis of the contexts associated with observable linguistic markers, namely, here, the *'alif-nûn* sequence, and aims at solving underlying ambiguity without resorting to deep morpho-syntactic analyses.

Key words :
Arabic NLP – linguistic ambiguity – context – Contextual Exploration Processing - linguistic marquers

Résumé

Une des grandes difficultés – au demeurant bien connue – du traitement automatique de la langue arabe est l'absence totale ou partielle des signes de « vocalisation » (ou de « voyellation »), et en contexte initial, de l'omission du signe notant la *hamza* dans les textes écrits, ce qui augmente considérablement les cas d'ambiguïté liés tant à l'homographie qu'à la polysémie lexicale. Ainsi par exemple, la séquence graphique *'alif-nûn* (ان), en l'absence de ce type de signes, peut correspondre à cinq mots-outils différents, relevant de différents champs grammaticaux. Dans cet article nous proposons une méthode de traitement automatique, l'Exploration Contextuelle (Desclés et al., 1991 ; Desclés 2006), basée sur l'analyse linguistique du contexte des marqueurs linguistiques observables, en l'occurrence ici la graphie *'alif-nûn,* afin de lever l'ambiguïté sous-jacente, sans avoir recours à des analyses morpho-syntaxiques profondes.

Mots clés

Traitement automatique de la langue arabe – ambiguïté linguistique – contexte – Exploration Contextuelle (EC) – marqueurs linguistiques.

Introduction :

Les premières traces de l'écriture arabe, sous la forme que nous connaissons aujourd'hui, remontent au IV^{ème} siècle de notre ère (toutefois, des inscriptions remontant au –IX^{ème} s. dans une écriture relevant de la famille sud-arabique ont été découvertes et analysées au cours des dernières décennies – Robin, 2001). Les premières graphies d'origine nabatéenne ou syriaque ne notaient pas les voyelles dans le corps du mot, fidèles en cela au modèle des langues sémitiques de l'Ouest. En ce qui concerne l'arabe, ce n'est qu'avec l'avènement des grandes campagnes de transcription des textes religieux au VIII^{ème} siècle que le besoin de fixer la prononciation sur l'écrit se fit ressentir (Abbott, 1939 ; Dichy, 1990). Plusieurs étapes se sont alors vraisemblablement succédées, selon une chronologie qui demeure malaisée à reconstituer, en raison du fait que les pratiques d'écriture n'intégraient que très partiellement les techniques de transcription tendant à la phonétisation de l'écrit : points diacritiques permettant de distinguer certaines consonnes, notation de la voyelle longue /â/ en milieu de mot au moyen de la lettre *'alif,* transcription, au moyen de signes diacritiques, des voyelles brèves et de certaines marques casuelles (dites de « nûnation », *tanwîn*), ajout de nouvelles graphies, etc.[1] Ce système aboutit aux différents signes de l'alphabet arabe actuel, qui compte :

- vingt-cinq consonnes, comportant, essentiellement, une forme finale (de mot) et une forme non-finale, les consonnes notées dans le corps du mot par un même graphisme étant distinguées par des points placés au-dessus ou au-dessous de la lettre, et correspondant à des *signes diacritiques primaires* (Dichy, 1990) :
  ب, ت , ث , ج , ح , خ , د , ذ , ر , ز , س , ش ,ص , ض , ط , ظ , ع , غ , ف ق , ك ل , م , ن , هـ
- deux segments-lettres (حرف) correspondant, soit à une consonne vocalique, soit à la voyelle longue homorganique : *wâw* (و), *yâ'* (ي) ;
- le segment-lettre *'alif* (ا), qui correspond, en début de mot à la consonne *hamza* (« coup de glotte »), et ailleurs, soit à la *hamza*, soit à la voyelle longue /â/ ;
- trois graphies supplémentaires : *hamza* (ء), notée, soit sous forme diacritique (plusieurs lettres pouvant lui servir de « support »), soit sur la ligne ; *'alif maqsûra* (ى) ; et *ta' marbûṭâ* (ة).

[1] Pour l'étude de la cohérence interne et la description linguistique de ce système graphique, voir Dichy, 1990.

Le système graphique de l'arabe inclut également dix autres *signes diacritiques secondaires*, traditionnellement désignés comme des « signes de vocalisation », ajoutés sous forme d'éléments suscrits ou souscrits :

- les trois voyelles brèves de l'arabe (d'où la désignation de l'ensemble des diacritiques secondaires du terme de vocalisation) : *fatḥa* (ـَ), *ḍamma* (ـُ), *kasra* (ـِ) ;
- sept autres signes : *sukun* (signe de quiescence ou absence de voyelle : ـْ), *šadda* (redoublement ou gémination d'une consonne : ـّ), les trois signes de *tanwin* (« nûnation », marques casuelles indéterminées : ـً, ـٌ, ـٍ), la *madda* (diacritique placé sur une lettre *'alif* et correspondant à une *hamza* suivie d'une voyelle longue *â* : آ), *wasla* (signe de la « *hamza* de liaison »), etc.

Dans la pratique, le code de l'écriture arabe penche vers la simplification : à part le cas des textes religieux ou didactiques, où la vocalisation est respectée, on s'aperçoit que partout ailleurs, seuls quelques signes de vocalisation sont maintenus à l'écrit, et de manière plutôt sporadique[2]. C'est donc l'absence de la totalité ou d'une partie de ces dix signes qui cause l'ambiguïté dans le traitement automatique ou même dans la lecture de textes par un humain[3]. Ainsi, si l'on ouvre n'importe quel journal de n'importe quel pays arabe, nous nous rendons à l'évidence que les mots des textes sont partiellement, ou, le plus souvent, complètement dépourvus de signes de « vocalisation », ou signes diacritiques secondaires (Dichy, 1990). Si le lecteur natif arabophone n'a souvent pas de difficulté à dégager le vrai sens des mots, ce type d'ambiguïté pose encore de réels problèmes à la machine. (Abbès et Dichy, 2008)

En faisant abstraction pour le moment d'autres difficultés liées au traitement automatique de l'arabe (absence de majuscules, agglutination[4] de certaines particules au début ou à la fin des mots, ordre relativement libre des mots dans la phrase, ponctuation non régulière, etc.), nous allons nous focaliser sur ce phénomène omniprésent et voir quels sont les moyens qui existent pour le traiter automatiquement.

Afin d'expliciter notre propos, nous avons choisi de traiter un cas d'ambiguïté assez fréquent dans les textes en arabe, celui de la graphie *'alif-nûn* (ان).

---

[2] Il serait fort intéressant d'ailleurs (cela sort du cadre de notre actuelle présentation) d'étudier la régularité d'apparition de ces signes dans les textes et de tenter de proposer un système de vocalisation automatique, non pas de toutes les lettres des mots d'un texte, mais uniquement des lettres qui portent ce genre de signes. Cette « vocalisation minimale et nécessaire », basée sur des règles linguistiques, peut résoudre beaucoup de cas d'ambiguïté dans le traitement automatique des textes en arabe. Pour un travail sur la vocalisation automatique entière ou partielle de l'arabe, voir (Ghénima, 1998).

[3] Certains mots, même en présence de signes de vocalisation, sont ambigus (عَلَّقَ / *accrocher* ou *commenter*). En leur enlevant ces signes, la combinatoire des possibilités de vocalisation augmente d'avantage.

[4] Leur nature cursive permet aux lettres en arabe de se lier les unes aux autres au sein d'un même mot (ex. : écoute / يسمع = ي+س+م+ع), mais aussi entre unité lexicale et morphèmes grammaticaux en position d'affixes ou de clitiques (Cohen, 1970 ; Desclés, dir., 1983 ; Dichy et Hassoun, eds., 1989 ; Dichy, 1997).

Les variantes de la graphie alif-nûn :

En l'absence de signes de vocalisation et de notation de la *hamza*, on peut trouver dans les textes en arabe une graphie *'alif-nûn* (ان) qui renvoie à l'un des cinq cas suivants[5]:

***'inna* (إنَّ)**

Le mot-outil *'inna* introduit une phrase nominale ou une phrase locative. Il est donc toujours suivi d'une phrase nominale dont le premier terme est un nom ou un pronom, sauf lorsque cette phrase indique une localisation, avec nom indéfini : إن في القصر جاسوسًا « Vraiment, [il y a] dans le palais un espion » (ou : « Il y a vraiment un espion au palais »).

Sa valeur est corroborative (elle « renforce » la valeur de vérité de l'assertion aux yeux de l'énonciateur)[6]. Sans correspondant à l'identique en français ce marqueur peut se traduire très imparfaitement par *il est vrai* ou *certes* (dont l'effet en français moderne inclut une concession à un point de vue opposé), ou de manière un peu moins approximative, en français parlé, par « vraiment »[7].

'Inna est, dans les textes médiévaux, le plus souvent placé en début de phrase :

"إِنَّنِي أَنَا اللَّهُ لا إِلَهَ إِلاَّ أَنَا فَاعْبُدْنِي وَأَقِمِ الصَّلٰوةَ لِذِكْرِي"

« Certes, c'est Moi Allah: point de divinité que Moi. Adore-Moi donc et accomplis la prière pour te souvenir de Moi. » (Coran, 20/14)[8]

Mais ce mot-outil peut aussi introduire une subordonnée complétive du verbe énonciatif قال / *dire* [9] :

قال ساركوزي إن " فرنسا تدعم بكل قواها مفاوضات السلام غير المباشرة بين سورية وإسرائيل عن طريق تركيا"

Sarkozy a dit que « la France soutient de toutes ses forces les négociations indirectes de paix entre la Syrie et Israël via la Turquie » (Syria News)

D'autres cas de figure moins fréquents peuvent être rencontrés ; il s'agit notamment de :

- *'inna* précédé par l'un des mots-outils suivants, sommairement décrits :
  - (ألا) *'alâ* et (أما) *'amâ*, mots-outils à valeur interpellative,
  - (كلا) *kallâ,* négation interpellative,
  - (حتى الابتدائية) *hatta* 'initiale de phrase', « et même (si) »,

[5] Nous ne prenons pas en compte dans cette énumération le verbe *'anna,* « geindre », « gémir », que l'on ne rencontre que dans les textes anciens, ou dans des jeux de devinette comme : انّ المريضُ كريمٌ .
(Solution : *'anna l-marîḍu karîmin,* « Le malade geignit comme une gazelle ».)

[6] Pour *'inna*, la valeur corroborative de la relation entre les deux termes de la phrase nominale a été identifiée par A. Roman (1990).

[7] Cette traduction reste approximative, les valeurs pragmatiques du corroboratif *'inna* ne pouvant toutes être rendues par un seul et même adverbe.

[8] Les traductions données ici visent simplement à rendre le sens général des exemples en arabe, et non pas à transposer les cas traités en français ou à en trouver des équivalents exacts.

[9] Il est intéressant de souligner qu'il s'agit du seul verbe acceptant une complétive introduite par *'inna*. Les autres verbes énonciatifs introduisant des phrases assertives font usage de *'anna.*

• l'expression d'un serment. Exemple :

Par Dieu, ton père, certes (vraiment), a raison / والله إن أباك لمحق

• (واو الحال) coordonnant *waw* introduisant une subordonnée circonstancielle. Exemple :

Je les ai rencontrés alors que (vraiment) j'étais malade / قابلتهم وإني لمريض

Dans les quatre premiers cas ci-dessus, *'inna* est précédé par un marqueur introduisant un énoncé ou modalisant celui-ci.

Notons que *'inna* peut être inclus dans un mot graphique incluant :

— des proclitiques, comme les coordonnants (و) *wâ-* ou (ف) *fa-* ;

— des enclitiques, comme les pronoms compléments (كما) *-kumâ*, (هم ـ) *-hum,* (ي ـ) *-î*, ou (نا) *-nâ*, etc.[10]

***'anna* (أنَّ)**

Ce mot-outil introduit une subordonnée complétive. Comme *'inna* cette dernière correspond à une phrase nominale dont le premier terme est un nom ou un pronom, sauf dans le cas où la proposition introduite par *'anna* indique une localisation, avec un nom indéfini : علمت أن في القصر جاسوسًا « Je sus (ou : j'ai appris) qu'[il y avait] dans le palais un espion ».

'Anna peut être précédé, notamment, de l'un des marqueurs suivants :

– un verbe, un nom ou une locution exprimant une constatation, une assertion, une estimation ou une information :

Il m'a été rapporté que la société avait fait faillite / بلغني أن الشركة قد أفلست

Ma conviction [est] que cette transaction [est] bénéciaire / اعتقادي أن هذه المعاملة رابحة

[Il est] clair que ce timbre-poste [est] précieux / من الواضح أن هذا الطابع قيّم

Je [suis] certain que les résultats étaient truqués / أنا متأكد أنّ النتائج كانت مزيفة

– un verbe au style indirect :

Elle a répondu que ses livres [sont] interdits en Egypte / أجابت أن كتبها محظورة في مصر

– un nom en position de « terme initial » (*mubtada'*) d'une phrase nominale[11] :

Un de tes défauts est que tu es négligent / من ذنوبك أنك مهمل

– un des marqueurs suivants : هذا و, ذلك, ذلك و, ربما, ليت, لعل, غير, لو, لولا, بما, إلا, أم, رغم كما, بيد, حتى, بحجة, ثم, يعني, من, هو,هي, في, الى...

Comme *'inna,* le mot-outil *'anna* peut être inclus dans un mot graphique comprenant :

[10] Sur la combinaison des mots-outils avec des morphèmes grammaticaux liés, voir, à ce même colloque, Dichy et Zmantar (2009).

[11] La phrase nominale en arabe commence habituellement par le *mubtada',* le « terme initial », défini, sur lequel on attire l'attention. La deuxième partie de la phrase est appelée *xabar*, « information » ou « attribut », qui décrit ou « qualifie » le terme initial.

— des proclitiques, comme le coordonnant (و) *wa-*, les prépositions (ل) *li-* ou (ب) *bi-,* ou le marqueur de comparaison (ك) *ka-,* exemple :

كأن مشيتها من بيت جارتها مرّ السحابة لا ريث ولا عجل

(vers du poète du VIIe s. al-ʾAʿšâ / (الأعشى

Sa démarche, sortant de la maison voisine,

fut comme un nuage passant, sans pluie, sans hâte...[12]

— des enclitiques, i.e. des pronoms compléments :

"قَالَتْ رَبِّ إِنِّي وَضَعْتُهَا أُنثَى"

« Elle dit, Seigneur, voilà que j'ai accouché d'elle, d'une fille » (Coran, 3/36*).*

***ʾan* (أنْ)**

Ce mot-outil est toujours suivi d'un verbe au subjonctif et précédé de l'un des éléments suivants :

— un marqueur exprimant la volonté, une intention, une obligation, une crainte ou une éventualité, qui peut être un verbe : طلب / *demander* ou خشي / *craindre*, un nom : المفروض / *[il y a] obligation*, un verbe impersonnel : يجب / *il faut*, une locution : لا بدّ / *[il est] nécessaire* (mot-à-mot : *pas d'échappatoire à*), etc. :

Je veux que tu écrives la lettre / أريد أن تكتب الرسالة

Le mieux est que tu partes tôt le matin / الأفضل أن تسافر في الصباح الباكر

— un verbe exprimant l'imminence :

La lune est sur le point de devenir pleine / أوشك القمر أن يصبح بدراً

– le mot-outil عسى / *peut-être* :

"وَعَسَى أَن تَكْرَهُواْ شَيْئًا وَهُوَ خَيْرٌ لَّكُمْ"

Il se peut que vous ayez de l'aversion pour une chose alors qu'elle vous est un bien.

(Coran, 2/216*)*

— l'un des mots-outils suivants : هو, هي - إما, أو- على, إلى - بلا, دون, قبل, بعد,منذ ...

Les voyageurs doivent se diriger vers le deuxième portail / على المسافرين أن يتوجهوا إلى البوابة الثانية

Mais *ʾan* peut aussi apparaître au début d'une phrase et constituer avec le verbe qui la suit le terme initial (*mubtadaʾ*) d'une phrase nominale : أن تأتي خير لك / *Le fait que tu viennes est mieux pour toi.*

[12] Le vers arabe dit « sortant de la maison de sa voisine » – ce qui, pour la traduction française, manquait de poésie.

Les morphèmes grammaticaux pouvant être inclus avec *'an* dans un mot graphique sont :

- comme pour *'inna* et *'anna,* des proclitiques comme (و / *wâ-*), des prépositions comme (ب / *-bî*, ل / *li-*) ou le marqueur de comparaison (ك / *ka-*) ;
- des la négation لا, *lâ*, dont le /l/ est amalgamé avec le /n/ de *'anna,* ce qui donne ألا / *'allâ*, *que ne... pas.*

_'in_ (إنْ) conditionnel

Ce mot-outil, partiellement semblable à *si* en français, marque la potentialité ou la probabilité d'une hypothèse. Il est presque toujours suivi d'un verbe.

"قُلْ إِن كُنتُمْ تُحِبُّونَ اللّهَ فَاتَّبِعُونِي"

*Dis : si vous aimez Dieu, alors suivez-moi. (Coran,*3/31*)*

Il apparaît le plus souvent au début de la phrase : إن تدرس تنجح / *Si tu étudies, tu réussiras*.

Un réemploi moderne lui permet d'introduire une complétive après un verbe de doute ou d'interrogation :

*Mon ami se demande si j'étais (*ou *si je suis) malade /* يتساءل صديقي إن كنت مريضاً

_'in_ (إنْ) négation

Il existe également un emploi de la forme *'in* – rare, mais toujours attesté dans les textes actuels – qui est d'une toute autre nature que le précédent : il ne s'agit plus du même morphème que celui de la conditionnelle. Le marqueur *'in* correspond à une négation, dans des constructions le plus souvent exceptives, exemple :

إن هو إلا صديق من أصدقائي

Ce n'est qu'un ami à moi (mot-à-mot : *Il n'est rien, sinon un de mes amis*)

Les morphèmes grammaticaux susceptibles d'être inclus avec les deux types de *'in* dans un mot graphique sont uniquement des proclitiques, par exemple, les coordonnants (و / *wa-*) ou (ف / *fa-*) :

"فإن انتهوا فإن الله غفور رحيم"

S'ils cessent, Dieu est, certes, Pardonneur et Miséricordieux. » (Coran, 2/192*)*

Synthèse

Nous avons vu que les quatre variantes vocalisées de la graphie *'alif-nûn* ont des fonctions différentes, et que chacun des mots-outils correspondants se lie, au sein du mot graphique, à des éléments linguistiques qui lui sont propres.

Dans ce qui suit, nous allons voir comment lever l'ambiguïté sur une occurrence non vocalisée et sans notation de la *hamza* de cette séquence. La méthode que nous utilisons, l'Exploration Contextuelle (Desclés et al., 1991 ; Desclés, 2006), nous offre les moyens nécessaires à l'analyse des contextes de marqueurs de surface et à la levée de l'éventuelle ambiguïté sous-jacente, à l'aide de connaissances linguistiques, légères à automatiser. Mais avant d'exposer notre approche, nous proposons de voir comment ce même phénomène a été traité dans le domaine du TAL.

Différentes approches de traitement

Une première approche pour traiter ce phénomène consiste à vocaliser les textes en associant à chaque unité lexicale l'ensemble de ses vocalisations potentielles et de résoudre, ou au moins de réduire, l'ambiguïté qui porte sur la vocalisation. Cette opération est en quelque sorte analogue à l'accentuation automatique de textes en français (Simard 1998) saisis en « typographie pauvre », c'est à dire sans les lettres accentuées. Exemples :

- cote → cote, coté, côte, côté.
- eleve → élève, élevé.
- precedent → précédent, précèdent.

La thèse de M. Ghenima (1998) portait précisément sur cette démarche, qui est d'une bien autre complexité en arabe que ce que l'on trouve en français. Pour Debili (1998), si pour le français, une entrée lexicale sans accents accepte en moyenne 1,34 accentuations potentielles, pour l'arabe, le taux monte à 11,54 (soit une ambigüité 8 à 9 supérieure). L'auteur avance l'exemple de la graphie (كتب/ *écrire*) pour laquelle il parvient à 21 vocalisations lexicales et casuelles différentes (ce chiffre, toutefois, ne correspond pas à nos propres estimations). D'autres analyses, basées sur l'expérience de l'analyseur morphologique de l'arabe mis au point par Xerox (Beesley, 2001) ou sur les analyseurs associés à la base de données DIINAR.1 (http://diinar.univ-lyon2.fr), ont présenté des données sur les problèmes d'ambigüité engendrés par la graphie de l'arabe (Abbès, Dichy et Hassoun, 2004 ; 2005).

Prenons un autre exemple : la graphie (علم). Hors contexte, ce mot peut recevoir sept vocalisations suivantes, qui correspondent à des variations du noyau du mot :

— عَلَم/ *‘alam (drapeau)* ; عِلْم/ *‘ilm (science) ;* عَلِمَ/ *‘alima (il a su) ;* عُلِمَ/ *‘ulima (il a été su) ;* عَلَّمَ/ *‘allama (il a enseigné)* ; عَلِّمْ/ *‘allim (enseigne* - impératif) ; عُلِّمَ/ *‘ullima (il a été enseigné).*

Il s’y ajoute, pour chacun des deux items nominaux ci-dessus (*‘alam, drapeau* et *‘ilm, science),* cinq variations casuelles.

Une autre approche est proposée par Audebert et Jaccarini (1986) dans leur analyse du cas de la graphie *’alif-nûn* parmi d'autres « tokens ». Même si le but des auteurs est le même que le nôtre ici (désambiguïser la graphie *’alif-nûn*), le moyen est tout à fait différent : les auteurs utilisent en effet, à l'aide d'automates finis[13], un analyseur morphologique sans dictionnaire, « guidé » par un analyseur syntaxique pour lever certaines ambiguïtés. Toutefois, comme les travaux sur la grammaire des formants du mot graphique en arabe l’ont fait peu à peu apparaître, malgré l’intérêt d’une telle approche pour la recherche, l’analyse du niveau du mot ne peut être mise en œuvre de manière performante sans une ressource lexicale qui associe aux entrées des spécificateurs « gérant » les relations des noyaux lexicaux avec les formants grammaticaux inclus dans la structure du mot (Dichy et Hassoun, 1989 ; Dichy, 1990, 1997 ; Dichy et Farghaly, 2007).

L'Exploration Contextuelle

L'approche que nous proposons ici est différente des autres types de travaux présentés ci-dessus, bien qu’elle ne soit pas incompatible avec une partie d’entre eux. Nous utilisons en effet la méthode de l'Exploration Contextuelle (EC), basée sur l'analyse du contexte des marqueurs observables dans les textes (Desclés *et al*. 1991). Ces marqueurs sont de véritables traces des relations grammaticales ou discursives exprimées par l'énonciateur, et relèvent de deux types : des marqueurs forts (indicateurs) qui indiquent, par le principe d'*abduction*, l'éventuelle présence de la relation dans le texte ; et des marqueurs complémentaires (indices) qui confirment ou infirment cette relation.

La modélisation d'une tâche avec l'EC s'appuie sur la notion de *point de vue de fouille* (pdv). Celle-ci consiste en une analyse d'un besoin spécifique de l'utilisateur, et correspond donc à une manière particulière de voir le texte : ce qui est valable dans l'analyse d'un pdv ne l'est pas forcément pour un autre. D'où l'émergence d'un autre type d'indices, les indices négatifs, dont la présence dans un contexte annule l'exécution des règles d'EC ; ceci aide ainsi à mieux délimiter les frontières d'un pdv par rapport à un autre. Ce mécanisme, qui distingue les indicateurs et les indices d'une part, et les indices positifs et les indices négatifs d'autre part,

[13] Voir à ce propos (Jaccarini, 1997) et (Desclés, 2006).

n'est pas possible avec la méthode proposée dans (Audebert et Jaccarini, 1986). Notons enfin que l'EC prend en charge la structure logique du texte (titre, sections, paragraphes et phrases).
Afin de mieux illustrer la démarche appliquée, nous allons traiter en détail le cas de la particule *'alif-nûn* utilisée selon un pdv relevant de la prise de parole, là où elle introduit des paroles rapportées. Il s'agit, en d'autres termes, du discours rapporté indirect (DR-I).

Stratégie de désambiguïsation

La modélisation de ce pdv consiste à distinguer entre les cas où la séquence *'alif-nûn* est employée dans un DR-I, et ceux qui correspondent à d'autres utilisations. Indicateurs, indices positifs et indices négatifs seront ainsi analysés dans ce sens au niveau de chaque segment textuel.

Nous considérons que les indicateurs du DR-I sont les particules *'alif-nûn* précédées par un indice énonciatif qui introduit le propos rapporté :

- *'inna* précédé du verbe قال / *dire*, ou de la locution بحسب / *selon* ;
- *'anna* précédé de verbes, comme أعلن / *annoncer*, de gérondifs, comme مؤكّداً / *en affirmant*, etc.
- *'an* précédé de verbes, comme أمر / *ordonner* ou طلب / *demander*, de noms énonciatifs, tels شرط / *condition*, etc.
- *'in* précédé de verbes, comme تساءل / *se demander*, de gérondifs, comme محلّفاً / *en faisant jurer*.

Tous ces indices sont pour nous des indices « positifs ». Leur présence confirme l'hypothèse déclenchée par la présence d'un indicateur dans le segment textuel considéré.

La suite du travail consiste à isoler tous les cas qui ne relèvent pas du DR-I, c'est à dire le reste des cas illustrés ci-dessus. Par conséquent, tout indice qui apparaît dans le contexte de la particule *'alif-nûn* dans une construction n'introduisant pas une prise de parole peut éventuellement constituer un indice « négatif » pour la recherche de constructions introduisant une parole. Exemples :

- *'in* ou *'inna* au début d'une phrase : ان تعمل تنجح / *si tu travailles tu réussiras* ; ان الجو حار / *vraiment, il fait chaud ;*
- certains morphèmes associés à *alif-nûn* au sein du mot graphique ; c'est le cas des proclitiques ف / *fa* et ك / *ka* ;
- des verbes exprimant la volonté : أريد ان / *je veux que* ;
- des verbes impersonnels يجب ان / *il faut que ;*
- des noms exprimant une assertion ou une constatation : من الأكيد ان / *il est certain que ;*

- des verbes exprimant l'imminence كاد ان / *être sur le point de ;*
- le marqueur عسى/ *il se peut que, peut-être ;*
- etc.

Tests sur corpus

Les indicateurs et indices collectés par le linguiste sont exploités par un système de règles linguistiques, l'ensemble étant manipulé par un moteur d'EC. L'implémentation informatique de cette méthode (Alrahabi, Ibrahim et Desclés, 2006) offre un traitement automatique en plusieurs étapes : conversion et dévocalisation du corpus, génération des formes fléchies de marqueurs linguistiques, segmentation et annotation du corpus.

En ce qui concerne le corpus, nous avons choisi des textes de nature hétérogène[14], et nous avons effectué des tests afin d'annoter uniquement les phrases dans lesquelles les graphies *'alif-nûn* ont un emploi d'introducteur de déclaration. Une première étape consiste pour nous à supprimer tous les signes de vocalisation qui existent dans le corpus. Cette étape est nécessaire car nous partons du postulat que les textes doivent être complètement dépourvus de signes de vocalisation[15]. L'étape de segmentation automatique a pour but de définir les espaces de recherche dans lesquels nous allons rechercher les marqueurs linguistiques (titres ou phrases). Dans notre cas, il s'agit de segments délimités par des signes forts de ponctuation comme le point. Ensuite, chaque espace de recherche peut être divisé en plusieurs autres espaces en fonction du nombre d'indicateurs trouvés dans la phrase[16].

L'étape d'annotation consiste, quant à elle, à identifier dans un premier temps les indicateurs. La présence de ceux-ci dans un espace de recherche va déclencher l'exécution des règles d'EC sous-jacentes et les prémisses de celles-ci vont être examinées. Si toutes les conditions sont satisfaites, l'annotation est attribuée à l'espace de recherche en question.

Voici un exemple de règle d'EC écrite dans un langage déclaratif :

Soit E l'espace de recherche suivant : toute phrase du texte

Rechercher les indicateurs de la liste "'alif-nûn" dans E

Si un indicateur 'alif-nûn existe dans E

Alors déclencher les règles sous-jacentes

Règle I

[14] Le corpus de tests est diversifié : des textes journalistiques (Al-Nahar, Tishreen, Al-Ahram, Al-Jazeera, Al-Sabah, Al-Alam, Al-Quds) et des textes littéraires (Al-Jaḥiz, IXe s., Yahyâ Haqqî, début du XXe s.) et religieux (*Coran* et hadith).

[15] Les formes de marqueurs linguistiques que nous recherchons dans les textes sont évidemment dépourvus de signes de vocalisation.

[16] À chaque fois qu'un indicateur est trouvé, le nouvel espace de recherche devient l'espace entre cet indicateur et la fin de la phrase.

Si l'indicateur n'est pas tout au début de E

Si l'indicateur est précédé d'un indice positif de la liste "indices-positifs-verbes" (ex. قال / *dire*)

Si entre l'indicateur et l'indice positif il n'existe pas d'indice négatif de la liste "indices-negatifs-verbes" (ex. أراد/ *vouloir*)

Alors annoter la phrase en cours avec l'annotation "DR-I"

Règle II

…

Ont été écrites, en l'étape actuelle du travail, 12 règles d'EC qui manipulent environ 550 marqueurs linguistiques (indicateurs, indices positifs et indices négatifs).

Discussion

Sur l'ensemble des 150 textes traités (179 262 mots), 163 phrases ont été annotées. Parmi ces résultats, 156 phrases sont correctement annotées (95,7%) et 7 phrases sont annotées alors qu'elles ne le devraient pas. Nous considérons que ce sont de bons résultats réalisés dans un temps de traitement relativement court[17]. Deux principales raisons sont à l'origine du bruit généré. La première est liée au repérage des indicateurs ou des indices à l'intérieur des guillemets de citation : ces marqueurs appartiennent à un plan énonciatif différent de celui des autres marqueurs de la règle. Exemple :

وقال موسى "مصلحتنا أن يستقر لبنان وألا تستمر الاغتيالات"

Moussa a dit « Notre intérêt est que le Liban se stabilise et que les assassinats ne se perpétuent pas »

La deuxième difficulté est caractérisée par des cas où un indicateur est repéré dans une relation différente de celle de l'indice positif, comme dans l'exemple :

وكان الوزير قد اعلن استقالته في وقت مسبق مع ان التهمة لم تكن قد ثبتت عليه بعد.

Le ministre avait déjà déclaré sa démission, alors que la présomption de sa culpabilité n'avait pas encore été confirmée.

Dans cette phrase nous avons un discours rapporté « nominalisé » (اعلن استقالته / *déclaré sa démission*), et un indicateur précédé d'un indice de pdv négatif (مع أن) qui n'avait pas été répertorié par nous.

La prochaine étape de ce travail portera sur un corpus plus large ; ceci nous permettra de faire un inventaire plus complet des indices, surtout négatifs, et de mesurer le « silence » dans les résultats, dont nous n'avons actuellement aucun indice.

[17] Le temps de traitement était de 91 secondes sur une machine Dell, avec MP Intel Core 2 Duo, et 2GO de mémoire.

Les valeurs sémantiques des différentes particules *'alif-nûn* peuvent aussi être organisées dans une carte sémantique, une sorte d' « ontologie linguistique » où les nœuds sont représentés dans les textes par les indicateurs et les indices.

Un des résultats de ce travail est la confirmation que la méthode utilisée (l'EC) offre une analyse linguistique fine basée uniquement sur des marqueurs de surface, sans avoir recours à des analyses automatiques lourdes dans le cadres des approches traditionnelles (considérant les analyses morphologiques, syntaxiques, etc. comme des étapes successives). Nous notons enfin que, le processus d'annotation étant complètement indépendant d'une langue donnée ou d'une tâche précise, le passage d'une langue à une autre ou bien le traitement d'une nouvelle tâche avec la méthode de l'EC est simple et méthodique, à condition d'avoir déjà construit bien évidemment les ressources linguistiques nécessaires, à savoir les marqueurs linguistiques et les règles d'EC.

Bibliographie

Abbès, Ramzi et Dichy, Joseph (2008) : « Extraction automatique de fréquences lexicales en arabe et analyse d'un corpus journalistique avec le logiciel AraConc et la base de connaissances DIINAR.1. » In : Heiden, Serge et Bénédicte Pincemain, *Actes des JADT 2008, 9es journées internationales d'analyse statistique des données textuelles (Proceedings of JADT 2008, 9th International Conference on Textual Data statistical Analysis)*, Lyon 12-14.03.2008, Presses Universitaires de Lyon, 2 vol., p. 31-44 – http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2008/pdf/abbes-dichy.pdf

Abbès, Ramzi, Dichy, Joseph et Hassoun, Mohamed (2004). « The Architecture of a Standard Arabic lexical database: some figures, ratios and categories from the DIINAR.1 source program. » In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*- COLING 2004 – University of Geneva, 28th August 2004 : 15-22.

– (2005). « Morpho-lexical ambiguities in the recognition of written Arabic word-forms, evidence from the DIINAR.1 lexical resource. » Colloque ACIDCA-ICMI'05 (International Conference on Machine Intelligence), Tozeur (Tunisie), 5-7 novembre 2005.

Abbott, Louisa N. (1939) : *The Rise of the North Arabic Script and its Kur'anic Development,* Chicago University Press.

Alrahabi, M. , Ibrahim, A.H., Desclés, J.-P. (2006). « Semantic Annotation of Reported Information in Arabic. » In: *FLAIRS 2006*, Florida, USA.

Audebert, C., Jaccarini, A. (1986). « À la recherche du HABAR, outils en vue de l'établissement d'un programme d'enseignement assisté par ordinateur. » *Annales islamologiques*, Tome XXII, Institut français d'archéologie orientale du Caire.

Beesley, Kenneth (2001). « Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. » In *ACL 39th Annual Meeting.* Workshop on *Arabic Language Processing; Status and Prospect,* Toulouse, 2001: 1-8.

Cohen, D. « Essai d'une analyse automatique de l'arabe ». In: David Cohen. *Études de linguistique sémitique et arabe.* Paris: Mouton, 1970, pp. 49-78.

Desclés, Jean-Pierre, dir. (1983). *Conception d'un synthétiseur et d'un analyseur morphologiques de l'arabe, en vue d'une utilisation en Enseignement assisté par Ordinateur* (H. Abaab, J.-P. Desclés, J. Dichy, D.E. Kouloughli, M.S. Ziadah), Rapport rédigé à la demande du Ministère des Affaires étrangères.

Desclés, J.-P., Jouis, C., Oh, H-G., Reppert D. (1991). « Exploration Contextuelle et sémantique: un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte. » In D. Herin-Aime, R. Dieng, J-P. Regourd, J.P. Angoujard (eds.), *Knowledge modeling andexpertise transfer*, pp.371--400, Amsterdam.

Desclés Jean-Pierre (2006). « Contextual Exploration Processing for Discourse Automatic Annotations of Texts. » *FLAIRS 2006*, Melbourne, Floride, 11-13 mai, Invited Speakers

Dichy, Joseph (1990). *L'Écriture dans la représentation de la langue : la lettre et le mot en arabe.* Thèse d'État, Université Lumière-Lyon 2.

– (1997). « Pour une lexicomatique de l'arabe : l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot. » *Meta 42*, printemps 1997, Québec, Presses de l'Université de Montréal : 291-306. www.erudit.org/revue/meta/1997/v42/n2/002564ar.pdf

Dichy, Joseph and Farghaly, Ali (2007). "Grammar-lexis relations in the computational morphology of Arabic". In Abdelhadi Soudi, Guenter Neumann and Antal Van den Bosch, eds., *Arabic Computational Morphology: Knowledge-based and Empirical Methods,* Dordrecht : Kluwer/Springer (series on Text, Speech, and Language Technology), chapter 7, p. 115-140.

Dichy, Joseph et Hassoun Mohamed, éds. (1989). *Simulation de modèles linguistiques et Enseignement Assisté par Ordinateur de l'arabe - Travaux SAMIA I.* Paris, Conseil International de la Langue Française.

Dichy, Joseph et Zmantar, Yasser (2009). « L'analyse automatique des mots-outils en arabe. » Actes de la *2ème Conférence internationale sur les Systèmes d'Information et l'Intelligence Economique (SIIE'2009)* 12,13 et 14 février 2009, Hammamet, Tunisie. http://www.siie.fr

Ghenima, Malek (1998). *Analyse morpho-syntaxique en vue de la voyellation assistée par ordinateur des textes écrits en arabe.* Thèse de doct., ENSSIB/Université Lyon 2.

Jaccarini, André (1997). *Grammaires modulaires de l'arabe. Modélisations, mise en œuvre informatique et stratégies*, thèse de doctorat, Université Paris IV-Sorbonne, 2 vol.

Robin, Christian Julien (2001). « Les inscriptions de l'Arabie antique et les études arabes. » *Arabica,* t. XLVIII, 2001, n°4 : 509-577

Roman, André (1990). *Grammaire de l'arabe,* Paris : P.U.F. (coll. « Que sais-je ? »).

Simard M. (1998). « Automatic insertion of accents in French text. » In *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing*, Grenade.