Analyse et désambiguïsation morphologiques de textes arabes non voyellés

Lamia Hadrich Belguith, Nouha Chaâben

Faculté des Sciences Économiques et de Gestion de Sfax – Laboratoire LARIS
l.belguith@fsegs.rnu.tn ; nouha.chaaben@laposte.net

Résumé

Dans ce papier nous proposons d'abord une méthode d'analyse et de désambiguïsation morphologiques de textes arabes non voyellés permettant de lever l'ambiguïté morphologique due à l'absence des marques de voyelles et aussi à l'irrégularité des formes dérivées de certains mots arabes (*e.g.* formes irrégulières du pluriel des noms et des adjectifs). Ensuite, nous présentons le système MORPH2, un analyseur morphologique de textes arabes non voyellés basé sur la méthode proposée. Ce système est évalué sur un livre scolaire et des articles de journaux. Les résultats obtenus son et très encourageants. En effet, les mesures de rappel et de précision globales sont respectivement de 69,77 % et 68,51 %.

Mots-clés : analyse morphologique, désambiguïsation morphologique, TALN arabe.

Abstract

In this paper we first propose a morphological disambiguation and analysis method of non voweled Arabic texts. This method could handle morphological ambiguities, caused by both the absence of the vowel marks and the irregularity of certain words (*e.g.* irregularity of the plural form for nouns and adjectives). Then, we present the MORPH2 system, a morphological analyser of non voweled Arabic texts, based on the proposed method. This system is evaluated on a school book and newspaper articles. The obtained results are very encouraging. Indeed, the global measures of recall and precision are 69,77 % and 68,51 % respectively.

Keywords: morphological analysis, morphological disambiguation, arabic NLP.

1. Introduction

Le Traitement Automatique du Langage Naturel (TALN) est un domaine à la fois scientifique et technologique en plein essor qui débouche sur des applications très diverses : correction automatique des erreurs, analyse de textes, génération automatique de résumés, extraction de connaissances, interrogation de bases de données en langage naturel, aide à la traduction, etc. En outre, la nécessité d'applications de TALN s'avère de plus en plus indispensable avec l'explosion d'Internet où le langage humain reste un vecteur d'information prépondérant. En particulier, pour l'arabe dont l'effectif de locuteurs dépasse les 200 millions. Comme la plupart des applications de TALN nécessitent une analyse morphologique du texte appréhendé, le développement de systèmes robustes traitant l'analyse morphologique s'avère indispensable. C'est dans ce contexte que se situe le présent travail[1].

Dans ce qui suit, nous commençons par une brève présentation de l'état de l'art sur l'analyse morphologique et nous citons quelques travaux relatifs à l'arabe. Ensuite, nous exposons

[1] Le présent travail entre dans le cadre d'un projet sur " La conception et le développement d'un système pour la détection et la correction des erreurs grammaticales dans des textes arabes" supporté par la Banque Islamique de Développement (BID) dans le cadre du programme YRSP.

quelques difficultés d'analyse morphologique de l'arabe. Après, nous proposons notre méthode d'analyse et de désambiguïsation morphologiques qui est implémentée à travers le système MORPH2 : un analyseur morphologique de textes arabes. Enfin, nous donnons les résultats d'évaluation de notre système MORPH2.

2. État de l'art sur l'analyse morphologique

L'analyse morphologique a fait l'objet de plusieurs travaux de recherche qui sont classés principalement en deux approches : soit que l'on utilise un lexique comportant tous les mots (sous leurs différentes formes possibles) avec leurs caractéristiques associées (Clavier, Lallich-Boidin, 1994), approche difficile à appliquer pour les langues qui ont une morphologie riche et complexe telle que la morphologie arabe ; soit que l'on réduise le lexique aux seules informations non calculables (*i.e.* formes canoniques, racines, etc.) et que l'on utilise des règles pour connaître le reste des informations (Belguith, 1999). La plupart des méthodes d'analyse morphologique de la langue arabe appartiennent à cette deuxième approche.

L'analyse morphologique est relativement avancée pour les langues latines. Ceci ne s'applique pas au cas de l'arabe pour diverses raisons, entre autres le manque de ressources linguistiques (*e.g.* corpus, lexiques de base, segmenteurs de textes en phrases). Pour cela, la majorité des travaux traitant la morphologie arabe se sont intéressées plutôt à l'étiquetage morphologique en se basant sur des méthodes d'apprentissage et une légère analyse morphologique (Khoja, 2001 ; El-Kareh et El-Ansary, 2000 ; Diab *et al.*, 2004 ; El Jihad et Yosfi, 2005). Cependant, on peut citer quelques systèmes d'analyse morphologique pour l'arabe tels que :

- Le système d'analyse morphologique de textes arabes (Tahir *et al.*, 2003) qui donne pour chaque mot une seule analyse morphologique, la première solution valide rencontrée. De ce fait, le système peut donner une solution non adéquate avec le contexte du mot.
- Le système d'analyse morphologique des noms arabes (Abuleil et Evens, 2004) permettant seulement la détermination de caractéristiques morphologiques des noms arabes.
- Le système Sebawai (Darwish, 2002) est un système d'analyse morphologique de surface utilisé dans une application de recherche d'information. Ce système s'intéresse seulement à la recherche des racines possibles d'un mot arabe donné.
- Le système AraParse (Ouersighni, 2002) est un système d'analyse morpho-syntaxique de l'arabe voyellé ou non voyellé utilisé pour la détection et le diagnostic des erreurs d'accord. Le système se compose de deux fonctions principales : une fonction d'analyse morpho-lexicale et la fonction de parsage.
- L'analyseur morphologique à états finis de Xerox (Beeseley, 2001) est un analyseur morphologique pour l'arabe utilisant les outils de Xerox de modélisation de langage à états finis. Il donne pour chaque mot toutes ses listes de caractéristiques morphologiques possibles.
- L'analyseur morphologique arabe basé-Web (Atwel *et al.*, 2004) est un analyseur morphologique traitant des textes arabes non voyellés dérivés de l'Internet. Il se base sur une méthode d'exploration contextuelle qui permet d'identifier le mot et ses caractéristiques contextuelles et donc de rechercher les affixes qui peuvent lui être associés.

3. Difficultés de l'analyse morphologique de l'arabe

3.1. Ambiguïtés dérivationnelles et flexionnelles

La flexion est la variation de la forme des mots en fonction de facteurs grammaticaux tel que la conjugaison pour les verbes (*exemple : le mot* " يتـأثــّرون " *(ils s'influencent) est le résultat de la concaténation du préfixe* "ي" *indiquant le présent et du suffixe* "ون " *indiquant le masculin pluriel du verbe* " تأثــّر "). Le problème en analyse morphologique de l'arabe se rapporte surtout au niveau de la dérivation qui est un phénomène plus complexe que la flexion. En effet, la dérivation est la formation de nouveaux mots à partir de mots existants. Dans le cas de la langue arabe, la plupart des mots sont dérivés à partir de racines trilitères ou quadrilitères. Le mot arabe n'est pas le résultat d'une simple concaténation de morphèmes comme c'est le cas pour l'anglais (*exemple : unfailingly = un+fail+ing+ly*), mais c'est à partir d'une racine, d'une combinaison de voyelles, de préfixes, d'infixes, de suffixes et d'un schème morphologique qu'on obtient un mot (*exemple : à partir de la racine* " أثــر " *(choisir/citer à) on peut dériver plusieurs verbes tel que* "تأثــّر" *(s'influencer) et plusieurs noms tel que* "متأثــّر"*(ému)*).

3.2. Ambiguïtés d'agglutination

Contrairement aux langues latines, en arabe, les articles, les prépositions, les pronoms, etc. collent aux adjectifs, noms, verbes et particules auxquels ils se rapportent. Comparé au français, un mot arabe peut parfois correspondre à une phrase française (Souissi, 1997) (*exemple : le mot arabe* "أتتذكّروننا" *correspond en français à la phrase* « *Est-ce que vous vous souvenez de nous* »). Cette caractéristique engendre une ambiguïté morphologique au cours de l'analyse. En effet, il n'est pas toujours facile de distinguer un proclitique ou enclitique d'un caractère original du mot. Par exemple, le caractère "و " dans le mot "وصل" *(il est arrivé)* est un caractère original alors que dans le mot " وفتح " *(et il a ouvert)*, il s'agit plutôt d'une proclitique.

3.3. Ambiguïtés dues à la non voyellation

La morphologie arabe est assez régulière lorsque les mots sont présentés sous leurs formes voyellées. Cependant, la majorité des documents arabes sont non voyellés sauf pour le Coran et pour certains ouvrages scolaires pour débutants et donc c'est pour cette raison que nous nous sommes intéressés à l'arabe non voyellé. En fait, les mots non voyellés engendrent beaucoup de cas ambigus au cours de l'analyse (*exemple : le mot non voyellé* " فصل " *pris hors contexte peut être un verbe au passé conjugué à la troisième personne du singulier* "فَصَلَ"*(il a licencié), ou un nom masculin singulier* " فَصْلٌ " *(chapitre/ saison), ou encore une concaténation de la conjonction de coordination* " فَ " *(puis) avec le verbe* " صل " *: impératif du verbe lier conjugué à la deuxième personne du singulier masculin*).

4. Notre méthode

La plupart des travaux existant sur l'analyse morphologique de l'arabe traitent le cas de l'arabe sous sa forme voyellée (voir § 2). De plus, la majorité des travaux qui s'intéressent à l'arabe non voyellé ne permettent pas de tenir compte des différents types d'ambiguïtés que nous avons présentés à la section 3 puisqu'ils proposent des méthodes d'analyse morphologique de surface (*i.e.* ces méthodes sont destinées à des applications qui nécessitent

seulement la connaissance de la base du mot tels que les systèmes de recherche d'information et de classification de documents). La méthode que nous proposons est une méthode d'analyse morphologique approfondie qui permet de déterminer pour chaque mot non seulement sa base ou sa racine mais aussi la liste de toutes ses caractéristiques morphologiques possibles, tout en tenant compte des différents types d'ambiguïtés cités précédemment (l'agglutination, l'affixation, la transformation, etc.). De plus, notre méthode permet l'analyse des différents types de mots arabes, à savoir, les noms, les verbes, les adjectifs, les particules, etc. Notre méthode d'analyse et de désambiguïsation morphologiques entre dans le cadre de l'approche computationnelle et repose sur cinq étapes (voir figure 1) à savoir, la segmentation du texte en mots, le prétraitement morphologique, l'analyse affixale, l'analyse morphologique et le post-traitement (Chaâben et Belguith, 2003).

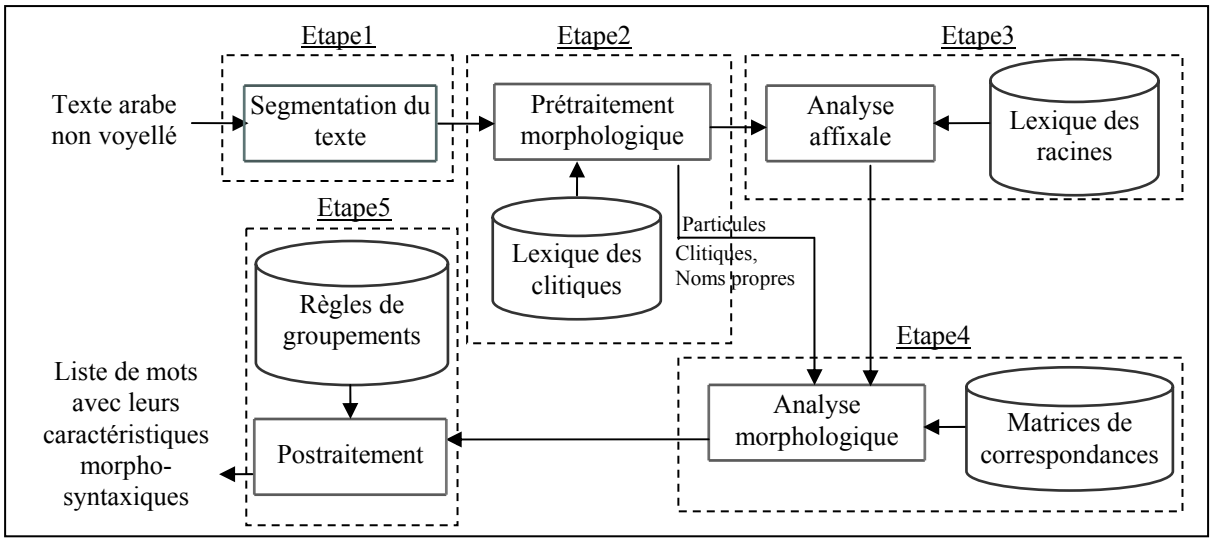

Figure 1. Les étapes de la méthode proposée

4.1. Segmentation du texte en phrases et en mots

La segmentation du texte en mots est faite en deux étapes : une segmentation du texte en phrases, en premier lieu, et une segmentation des phrases en mots, en second lieu. La segmentation du texte en mots est réalisé par le système STAr (Belguith *et al.*, 2005), un segmenteur de textes arabes basé sur l'exploration contextuelle des signes de ponctuation, des mots connecteurs jouant le rôle de séparateurs de phrases ainsi que celles de certaines particules, telles que les conjonctions de coordination. La segmentation de la phrase en mots se base sur la détection des espaces, des signes de ponctuation et de certains caractères spéciaux.

4.2. Prétraitement morphologique

Cette étape consiste à supprimer les proclitiques et les enclitiques pouvant être agglutinés au mot en se basant sur les listes de proclitiques et d'enclitiques. Le mot restant sera filtré pour tester s'il s'agit d'une particule, d'un nom propre, d'un nombre ou d'une date. Si ce n'est pas le cas, il va subir l'analyse affixale dans le but de déterminer toutes ses caractéristiques morphologiques possibles.

4.3. Analyse affixale

La langue arabe possède des propriétés structurelles spécifiques. Un mot du vocabulaire est constitué d'une racine (trilitère ou quadrilitère) ou d'une forme canonique à laquelle on ajoute une combinaison affixale formée d'un préfixe, d'un infixe ou deux, et d'un suffixe. L'analyse affixale a pour objectif de reconnaître les éléments de base qui entrent dans la constitution d'un mot à savoir, la racine (R) ou la forme canonique et les affixes (préfixe (P), infixe (I) et suffixe(S)). Cette opération est effectuée en plusieurs étapes (Ben Hamadou, 1993). D'une étape à l'autre un mécanisme de filtrage permet d'éliminer les décompositions parasites reconnues. On distingue les principales étapes suivantes : identification des couples (P, S), identification des triades affixales candidates, filtrage lexical, contrôle des associations (R) et (P, I, S) et reconnaissance des transformations.

4.4. Analyse morphologique

Cette étape consiste à déterminer, à partir de la forme (R, P, I, S) obtenue pour chaque mot, toutes ses caractéristiques morpho-syntaxiques possibles (*i.e.* partie de discours, genre, nombre, temps, personne, etc.). La détection des caractéristiques morpho-syntaxiques se fait en trois phases (Belguith, 1999). La première phase consiste à identifier la catégorie principale du mot (partie de discours), à savoir verbe, nom, adjectif, etc. La deuxième phase permet de déterminer pour chaque catégorie, identifiée au niveau de la phase 1, la liste de ses caractéristiques morphologiques. Un filtrage des listes de caractéristiques s'effectue en une troisième phase.

4.5. Post-traitement

Cette étape consiste à identifier les groupements de mots qui peuvent exister dans une phrase. Il s'agit de vérifier pour chaque couple de mots successifs ayant la catégorie NOM, s'il représente un groupement de mots ou un cas d'annexion et ce en se basant respectivement sur un lexique de groupement de mots et un ensemble de règles d'annexion. En cas de détection d'un groupement de mot ou d'une annexion, ces mots en question seront considérés comme étant une seule unité morphologique. Notons que l'annexion prend les caractéristiques morpho-syntaxiques de l'annexé (مضاف). Celles du complément du nom (مضاف إليه) seront stockées car elles seront utiles pour la détection des annexions complexes telles que les annexions composées d'un nom, d'un nom déterminé et d'un adjectif (*exemple :* قاعة المعمل الكبير *(La salle de la grande usine)*) ou celles composées d'un nom et d'un groupement de coordination (*exemple :* إنتاج البذور والزهور *(La production des graines et des fleurs)*).

5. Exemple illustratif

Dans ce qui suit, nous appliquons les différentes étapes d'analyse sur le mot "المصطلحات" (*les termes*). L'étape de prétraitement morphologique permet de décomposer le mot en question en un enclitique "ال" *(les)* et le mot "مصطلحات" *(termes)*. Le mot "مصطلحات" subit ensuite l'analyse affixale qui associe à chaque mot la liste des triades affixales et des racines valides. On obtient donc la décomposition suivante : (préfixe = "م", infixe ="ت", suffixe ="ات", racine= " صلح"). L'infixe "ت", qui occupe la troisième position dans le mot est déduit suite à l'application d'une règle de transformation. En effet, puisque la première lettre de la racine est "ص", l'infixe "ت" est transformé en "ط". Enfin, l'étape d'analyse morphologique consiste à associer à la triade affixale toutes ses caractéristiques morphologiques possibles, et dans ce cas la seule liste valide est : nom, pluriel masculin déterminé non humain.

6. Le système MORPH2

Nous présentons dans cette section notre système MORPH2, un analyseur morphologique basé sur la méthode proposée et permettant l'analyse morphologique de textes arabes non voyellés. Ce système est réalisé avec le langage de programmation JAVA et utilise un lexique XML (Chaâben et Belguith, 2004).

Figure 2. Résultat d'analyse de la phrase "استمتع الأولاد بالجولة"
(Les enfants se sont amusés par la promenade)

MORPH2 utilise dans chaque étape d'analyse un ensemble de données nécessaires au traitement : dans l'étape de prétraitement morphologique, il s'agit du lexique des proclitiques, enclitiques, particules ; dans l'étape d'analyse affixale du lexique des triades affixales et des racines (3 266 racines trilitères et quadrilitères) ; dans l'étape d'analyse morphologique enfin du lexique des noms dérivés et primitifs et le lexique des correspondances entre forme canoniques et formes dérivées. La figure 2 montre le résultat de l'analyse de la phrase " استمتع الأولاد بالجولة" (les enfants se sont amusés par la promenade). Ainsi, pour chaque mot, on obtient ses différentes analyses morphologiques possibles. Le mot "استمتع " peut être un verbe au passé conjugué à la troisième personne du singulier masculin à la voix active "اسْتَمْتَعَ " (est amusé) à la voix passive "اسْتُمْتِعَ " (s'est amusé). Le deuxième mot " الأولاد" (les enfants) est un nom dérivé, masculin pluriel, humain et déterminé. Et, le mot " بالجولة " (par la promenade) est un nom dérivé, nom singulier féminin auquel est agglutinée la proclitique "بـ" (par) qui est une préposition. Le résultat de l'analyse est enregistré dans un fichier XML.

7. Évaluation de MORPH2

Pour évaluer MORPH2, nous avons utilisé deux corpus. Le premier corpus est un livre scolaire tunisien utilisé dans l'enseignement de base. Il est composé de 81 textes arabes non voyellés contenant 29 188 mots. Les textes de ce corpus portent sur des thèmes différents (vie sociale, culture, découvertes, etc.) et sont généralement des textes de type narratif comportant plus de verbes que de noms. On trouve aussi des textes sur les thèmes de la nature, des sentiments, etc. qui sont généralement des textes descriptifs comportant plus de noms et d'adjectifs, que de verbes. Quant au deuxième corpus, il est extrait à partir du web et représente un ensemble d'articles de journaux de divers thèmes contenant 22 216 mots. L'évaluation consiste à calculer pour chaque catégorie de mots (*i.e.* verbe, nom, particule,

nom propre) ses mesures de « rappel » et de « précision ». Ensuite, nous déterminons ces mesures pour les deux corpus d'évaluation. Le tableau 1 présente la répartition par catégorie, des mots dans les corpus d'évaluation.

Corpus	Mots différents	Verbes	Noms	Particules	Noms propres
Livre de 8ième de base	14 151	5 878	7 552	548	447
Articles de journaux	8 970	2 022	6 552	494	377

Tableau 1. Répartition des mots (en catégories) dans les corpus d'évaluation

Le tableau 2 donne les résultats de l'évaluation de MORPH2. D'après ce tableau, nous constatons que l'analyseur a pu analyser avec succès presque 72 % des verbes du premier corpus et 67% des verbes du deuxième corpus. Les cas d'échec correspondent essentiellement aux verbes comportant des voyelles longues. À ces cas s'ajoutent les verbes hamzés et les verbes de radicales redondantes (*i.e.* verbes dérivant de racines qui ont la deuxième et la troisième radicales identiques, *exemple :* "عدد" *(compter)*). Pour les verbes faibles (verbes comportant des voyelles longues) le problème réside dans le fait que les voyelles longues des verbes peuvent être transformées ou même omises lors de la conjugaison. Les cas d'échecs pour les noms sont dus essentiellement à l'absence de la forme canonique du nom dans le lexique et aux transformations des voyelles longues et de la lettre hamza. Nous remarquons, en plus, que le nombre de solutions possibles par nom est de 1,26 (voir tableau 2) c'est-à-dire que l'analyse de la plupart des noms est non ambiguë (une seule liste de caractéristiques est générée). Les noms qui ont plus d'une analyse possible sont ceux ayant le suffixe "ين" qui indique que le nombre du nom est soit le pluriel soit le duel. Pour les particules telles que les prépositions, les interrogatifs, les adverbes, etc., les cas d'échec sont dus principalement au problème d'agglutination. De plus, on remarque que l'analyse des particules (1,54 solutions possibles par particule) est plus ambiguë que celle des noms (1,26 solutions possibles par nom). Ceci s'explique par le fait qu'une particule peut avoir plusieurs caractéristiques morphologiques (*exemple : la particule* "من" *peut être une préposition* "مِنَ" *(depuis) ou un interrogatif* "مَنْ" *(qui)*).

Corpus	Verbes		Noms		Particules		Noms propres		Mesures globales	
	R	P	R	P	R	P	R	P	R	P
Corpus 1	**78,01 %**	**71,26 %**	**52,35 %**	**51,73 %**	**72,57 %**	**70,02 %**	**73,71 %**	**73,71 %**	**69,77 %**	**68,51 %**
Corpus 2	**72,61 %**	**66,48 %**	**40,11 %**	**40,55 %**	**71,50 %**	**70,67 %**	**60,71 %**	**60,71 %**	**50,68 %**	**50,88 %**
Nombre de solutions possibles[2]	**3,45**		**1,26**		**1,54**		**1**		**2,7**	

Tableau 2. Résultas de l'évaluation du système MORPH2

[2] On désigne par nombre de solutions possibles d'un mot le nombre de solutions (listes de caractéristiques morpho-syntaxiques) pertinentes résultant de l'analyse morphologique de ce mot.

L'évaluation globale consiste à calculer pour chaque mot ses mesures globales de rappel et de précision en tenant compte de toutes les décompositions possibles du mot. Cette évaluation (voir tableau 2) montre une large différence entre les mesures concernant les deux corpus. Cette différence est due aux problèmes déjà décrits pour chaque catégorie de mot surtout ceux qui concernent les noms. Les cas d'échec, comme déjà présenté sont dus généralement aux transformations des voyelles longues et du hamza dans les mots ou à l'absence de la racine ou de la forme canonique du mot dans notre lexique. Pour le deuxième corpus (articles de journaux), les fautes d'orthographe, les omissions de « chadda » et la non présentation du hamza sont les principales causes de l'écart avec les mesures de rappel et de précision du premier corpus.

8. Conclusion

Nous avons proposé dans cet article une méthode pour l'analyse et la désambiguïsation morphologique de textes arabes non voyellés. Cette méthode entre dans le cadre de l'approche computationnelle et permet une analyse morphologique approfondie. Elle se base sur un lexique réduit et permet par des calculs, de déterminer pour chaque mot ses différentes caractéristiques morpho-syntaxiques (*i.e.* partie de discours, genre, nombre, temps, etc.), en tenant compte des différents cas d'ambiguïté tel que l'annexion, l'agglutination, etc.

Nous avons aussi présenté notre système MORPH2 basé sur cette méthode. Ce système est actuellement intégré dans deux systèmes d'analyse de l'arabe: MASPAR et DECORA. Le système MASPAR (Multi Agent System for Parsing Arabic) est un système multi-agent pour l'analyse syntaxique des textes arabes non voyellés (Aloulou *et al.*, 2003). Il utilise notre analyseur morphologique en tant qu'agent collaborant avec ses autres agents (*i.e.* agent « Segmenteur », agent « Syntaxe », agent « Ellipse » et agent « Anaphore »). L'agent « Segmenteur » représente le système STAr (Belguith *et al.*, 2005) permettant la segmentation du texte en phrases. L'agent « Syntaxe » reçoit de l'agent « Morphologie » une liste contenant toutes les solutions morphologiques possibles pour un mot. L'objectif de l'agent « Syntaxe » est de trouver les bonnes analyses syntaxiques d'une phrase tout en utilisant une grammaire des schémas de règles HPSG (Head-driven Phrase Structure Grammar) et en exploitant fortement le concept d'unification (Bahou *et al.*, 2006). MORPH2 est intégré aussi dans le système DECORA (DEtection CORrection des Accords) de détection et de correction des erreurs d'accord dans les phrases écrites en arabe non voyellé (Belguith et Ben Hamadou, 2004). MORPH2 fournit toutes les caractéristiques morpho-syntaxiques possibles des mots d'une phrase au système DECORA pour qu'il puisse détecter les erreurs d'accord.

Comme perspective, nous envisageons d'abord de comparer notre méthode à une méthode ligne de base. Cette méthode consiste à déduire la catégorie du mot à partir de son préfixe et/ou de son suffixe (*exemple : déduire qu'il s'agit d'un nom lorsque le mot se termine pas "ـة" ou commence par "ال"*). Ensuite, nous visons à étendre les lexiques utilisés afin de couvrir le maximum de racines et de formes canoniques de l'arabe. Enfin, nous envisageons d'ajouter une étape de désambiguïsation permettant de prendre en compte le contexte du mot pour réduire le nombre de solutions possibles.

Références

ABULEIL S., EVENS M. (2004). « Classify Arabic Nouns for Morphology Systems ». In *Actes de TALN 2004*.

ALOULOU C., HAMMAMI MEZGHANI S., BELGUITH HADRICH L., HADJ KACEM A. (2003). « Implémentation du système MASPAR selon une approche multi-agent ». In *Actes du 8e Colloque International sur les Techniques d'Analyse Syntaxique*.

ATWEL E., AL-SULAITI L., AL-OSAIMI S., ABU AHAWAR B. (2004). « A Review of Arabic Corpus Analysis Tools ». In *Actes de TALN 2004*.

BAHOU Y., BELGUITH HADRICH L., ALOULOU C., BEN HAMADOU A. (2006). « Adaptation et implémentation des grammaires HPSG pour l'analyse de textes arabes non voyellés ». In *Actes du 15e congrès francophone AFRIF-AFIA Reconnaissance des Formes et Intelligence Artificielle (RFIA'06)*.

BEESLEY K. (2001). « Finate-State Morphological Analysis and Generation of Arabic at Xerox Research : Status and Plans in 2001 ». In *Actes du Arabic NLP Workshop at ACL/EACL 2001*.

BELGUITH HADRICH L, BACCOUR L., MOURAD G. (2005). « Segmentation des textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules ». In *Actes de TALN 2005* : 451-456.

BELGUITH HADRICH L., BEN HAMADOU A. (2004). « Traitement des erreurs d'accord : une analyse syntagmatique pour la vérification et une analyse multicritère pour la correction ». In *Revue d'Intelligence Artificielle (RSTI- RIA)* 18 (5/6) : 679-707.

BELGUITH HADRICH L. (1999). *Traitement des erreurs d'accord de l'arabe basé sur une analyse syntagmatique étendue pour la vérification et une analyse multicritères pour la correction.* Thèse de doctorat en informatique, Faculté des Sciences de Tunis.

BEN HAMADOU A. (1993). *Vérification et correction automatique par analyse affixale des textes écrits en langage naturel : le cas de l'arabe non voyellé.* Thèse d'État en informatique, Faculté des sciences de Tunis.

CHAÂBEN N., BELGUITH HADRICH L. (2004). « Implémentation du système MORPH2 d'analyse morphologique pour l'arabe non voyellé ». In *Actes des Quatrièmes journées scientifiques des jeunes chercheurs en Géni-Electrique et Informatique (GEI'2004)*.

CHAÂBEN N., BELGUITH HADRICH L. (2003). « L'étiquetage morpho-syntaxique : comment lever l'ambiguïté dans les textes arabes non voyellés ? ». In *Actes des Troisièmes journées scientifiques des jeunes chercheurs en Génie Électrique et Informatique (GEI'2003)* : 41-44.

CLAVIER V., LALLICH-BOIDIN G. (1994). « Modélisation linguistique de la suffixation en vue de l'analyse automatique ». In *Revue TAL* 35 (2) : 129-144.

DARWISH K. (2002). « Building a Shallow Arabic Morphological Analyzer in One Day ». In *Actes du workshop Computational approaches to Semitic languages* : 47-54.

DIAB M., HACIOGLU K., JURAFSKY D. (2004). « Automatic Tagging of Arabic text : from raw text to base phrase chunks ». In *Proceedings of HLT/NAACL-200* : 149-152.

EL JIHAD A., YOUSFI A. (2005). « Étiquetage morpho-syntaxique des textes arabes par modèle de Markov caché ». In *Actes de RECITAL 2005* : 649-654.

EL-KAREH S. AL-ANSARY S. (2000). « An Arabic interactive Multi-feature POS tagger ». In *Actes de ACIDCA'2000* : 83-88.

KHOJA SH. (2001). « APT: Arabic part-of-speech tagger ». In *Proceedings of The Student Workshop at the second meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001)* : 20-26.

OUERSIGHNI R. (2002). « L'analyse morpho-syntaxique de l'arabe voyellé ou non voyellé : le système AraParse ». In *Actes de l'assemblée internationale du traitement automatique de la langue arabe*.

SOUISSI E. (1997). *Étiquetage grammatical de l'arabe voyellé ou non.* Thèse de doctorat, Université de Paris III.

TAHIR Y., CHENFOUR N., HARTI M. (2003). « Realization of a morphological analyzer for Arabic language text ». In *Proceedings of American workshop on the information technology*.