Analyse lexicale et morphologique de l'arabe standard utilisant la plateforme linguistique NooJ

Slim Mesfar

Université de Franche-Comté – LASELDI
mesfarslim@yahoo.fr

Résumé

Cet article décrit un système de construction du lexique et d'analyse morphologique pour l'arabe standard. Ce système profite des apports des modèles à états finis au sein de l'environnement linguistique de développement NooJ pour traiter aussi bien les textes voyellés que les textes partiellement ou non voyellés. Il se base sur une analyse morphologique faisant appel à des règles grammaticales à large couverture.

Mots-clés : TALN, NooJ, langue arabe, analyse lexicale, analyse morphologique, grammaire morphologique, agglutination, voyellation.

Abstract

This article describes the construction of a lexicon and a morphological description for standard Arabic. This system uses finite state technology, within the linguistic developmental environment NooJ, to parse vowelled texts, as well as partially vowelled and unvowelled ones. It is based on large-coverage morphological grammars covering all grammatical rules.

Keywords: NLP, NooJ, arabic language, lexical analysis, morphological analysis, morphological grammar, agglutination, vocalisation.

1. Introduction

La langue arabe, quatrième langue au monde[1], connaît un accroissement des contenus textuels, surtout en ligne, avec plus de 20 000 sites arabes sur le Web et plus de 300 millions utilisateurs. À ce jour, le traitement et l'exploitation de ces ressources documentaires présentent encore un défi pour les chercheurs dans le domaine du traitement automatique des langues naturelles. Dans le cadre de nos recherches, nous avons entrepris la construction d'un module pour l'arabe au sein de la plateforme linguistique de développement NooJ ; notre but est d'implémenter une composante d'analyse automatique des textes écrits en arabe standard. Cette composante servira à mieux comprendre la langue partant d'une description de son vocabulaire et de sa syntaxe transformationnelle selon la théorie de Harris et de Chomsky.

2. État de l'art

Depuis le début des années 60, et à partir du premier essai d'analyse automatique proposé par David Cohen, l'un des premiers théoriciens du domaine du TAL, des recherches sont poursuivies dans le cadre du traitement automatique de la langue arabe. En 1983, partant d'une analyse morphologique minimaliste basée sur le principe que toute forme linguistique arabe se traduit en schème et racine, les recherches se sont développées pour arriver à la construction du premier analyseur morphologique à deux niveaux de l'arabe (Koskenniemi,

[1] Source: *Ethnologue*, 13e édition, Barbara F. Grimes Editor, École d'été en Linguistique, 1996.

1983), repris et remanié dans le cadre du projet ALPNET utilisant des automates finis permettant uniquement la concaténation de morphèmes (Beesley et Buckwalter, 1989). Dès 1996, l'équipe de recherche de Xerox a enrichi ce même système par un algorithme de combinaison automatique entre racines et schèmes ; ces travaux se basent sur les dictionnaires du projet ALPNET qui ont été considérablement modernisés, utilisant les transducteurs à états finis de Xerox (Beesley, 2001). L'adéquation de cette technologie au traitement automatique des langues naturelles est bien connue, nous l'utilisons aussi avec l'environnement de développement NooJ, que nous décrivons ci-après.

3. Description de la langue arabe

La langue arabe est une langue sémitique présentant deux grandes caractéristiques faisant le sujet de nombreux travaux de recherche : l'agglutination et la non-vocalisation. En effet, chaque forme[2] d'un écrit en arabe peut correspondre à une suite d'un ou plusieurs préfixes, un radical et un ou plusieurs suffixes. Les radicaux sont elles-mêmes des formes fléchies et dérivées à partir de lemmes.

La non-vocalisation due à une absence des voyelles brèves dans les textes courants entraîne un haut degré d'ambiguïté. Si elles sont présentes, les voyelles brèves sont représentées par des diacritiques qui apparaissent au-dessus ou en dessous des consonnes qu'elles suivent. En principe, seuls le Coran et les livres d'enseignement sont vocalisés ; le traitement automatique de l'arabe doit pouvoir traiter des textes vocalisés mais aussi des textes non vocalisés.

Pour remédier à ces deux problèmes, nous avons utilisé des automates à états finis que nous avons associés à des dictionnaires de lemmes.

4. NooJ et le traitement de l'arabe

NooJ[3] est un environnement linguistique de développement qui peut analyser des corpus importants en temps réel. Il inclut des outils pour construire, tester et maintenir des descriptions formalisées à large couverture des langues naturelles (sous forme de dictionnaires et de grammaires électroniques). Les dictionnaires et les grammaires sont appliqués aux textes afin de localiser les modèles morphologiques, lexicologiques et syntaxiques, enlever des ambiguïtés, et étiqueter des mots composés et simples (Silberztein, 2005). NooJ peut construire des concordances lemmatisées de grands textes à l'aide de grammaires à états finis et algébriques, et peut aussi effectuer des opérations de transformation sur des textes en cascade, afin de les annoter, ou produire des paraphrases.

Le module lexical de NooJ, utilisé tout au long de cet article, se base sur des opérateurs de transformations à l'intérieur des formes et des graphes morphologiques décrivant des règles grammaticales à large couverture. Bien que certains opérateurs de transformations soient prédéfinis dans NooJ (*e.g.* <L> : touche de déplacement vers la gauche, <R> : touche de déplacement vers la droite, <S> : Suppression du caractère courant, etc.), nous pouvons les redéfinir ou en ajouter quelques-uns. Ces transformations fonctionnent sur une pile, elles nécessitent un temps de transformation en O(n). Ainsi, elles garantissent une correspondance entre le lemme et la forme fléchie correspondante en un temps linéaire. Quant aux grammaires morphologiques, elles sont construites en utilisant l'éditeur de graphes de NooJ et représentées sous forme de transducteurs à états finis (FST). Elles représentent des séquences

[2] Une forme est une suite de graphèmes se trouvant entre deux blancs ou ponctuations dans un texte.

[3] Le téléchargement libre et le manuel d'utilisation de la plateforme linguistique NooJ sont disponibles à l'adresse : http://www.nooj4nlp.net

de lettres et associent leurs reconnaissances à la production des informations lexicales correspondantes (étiquette grammaticale, un ensemble d'informations sémantiques, etc.).

5. Construction du lexique arabe

Étant donné que toute analyse linguistique doit passer par une première phase d'analyse lexicale, qui consiste à tester l'appartenance de chaque mot du texte au vocabulaire de la langue (Revuz, 1991), nous commençons notre travail par une phase de formalisation du vocabulaire de l'arabe. Ce travail a commencé par la formalisation de trois ensembles : les verbes, les noms et les particules.

5.1. Les verbes

Le dictionnaire des verbes contient 10 000 entrées[4] complètement voyellées. Chaque verbe, ramené à la 3e personne du singulier à l'accompli actif, est associé à un modèle de flexion (parmi 130 modèles développés pour la totalité des verbes) (Abou Il Azm, 2003).

Par modèle de flexion nous désignons l'ensemble des transformations permettant d'obtenir, à partir d'une entrée lexicale, l'ensemble de ses formes conjuguées. Ces paradigmes flexionnels incluent le mode (indicatif, subjonctif, apocopé et impératif), la voix (active et passive), le genre et le nombre, ce qui donne, en moyenne, 122 formes fléchies par entrée lexicale.

Exemple : 'كَلَّمَ', V+Tr+FLX[5] = V_kallama (kallama – *parler à quelqu'un*)[6]

Parmi les 122 transformations flexionnelles qui sont intégrées dans le paradigme flexionnel V_kallama, en voici une : « <LW>*yu*<R4><S>*i*<R><S>*u*/A+P+3+m+s ». Cette transformation NooJ signifie : positionner le curseur (|), initialement placé à la fin du mot (*kallama*|), à la tête du lemme par un déplacement vers la gauche (<LW>[7]) (|*kallama*), insérer (*yu*) (*yu*|*kallama*), sauter quatre lettres vers la droite (<R4>) (*yukall*|*ama*), effacer la lettre suivante (<S>) (*yukall*|*ma*), insérer la voyelle (*i*) (*yukalli*|*ma*), sauter une lettre vers la droite (<R>) (*yukallim*|*a*), effacer la lettre suivante (<S>) (*yukallim*|) et enfin insérer la voyelle finale (*u*) (*yukallimu*|).

Cette opération permet de générer la forme suivante : 'يُكَلِّمُ' (*yukallimu* – il parle à qq'un) qui sera associée aux informations flexionnelles : V+Tr+A+P+3+m+s, *i.e.* verbe transitif direct (V+Tr) conjugué au masculin (m) singulier (s), troisième personne (3), présent de l'indicatif (P) et voix active (A).

5.2. Les noms

Les noms sont décrits de trois façons différentes :

(1) Nous avons construit un dictionnaire qui contient environ 15 000 entrées sous forme de noms primitifs[8]; par exemple, le nom 'كُرْسِيّ' (*korsiyy* – chaise). Chaque entrée est ramenée au masculin singulier.

[4] La liste des verbes a été construite par Ibtihal Farawi et Slim Mesfar lors de leurs travaux de recherche sur l'arabe au LASELDI.

[5] FLX : fonctionnalité permettant la description des formes fléchies potentielles à partir d'un lemme.

[6] Nous associons à chaque forme écrite en arabe, délimitée par des apostrophes, sa transcription en caractères latins – en italique – et sa traduction en français.

[7] Le sens des déplacements tient compte de l'inversion de l'orientation droite-gauche des mots écrits en arabe pour une orientation gauche-droite dans les formes translittérées.

[8] Un nom primitif désigne un nom qui ne peut pas être dérivé d'un verbe.

(2) Nous avons associé les verbes décrits ci-dessus à des descriptions morphologiques pour représenter l'ensemble des déverbaux (*i.e.* des noms qui proviennent de verbes). Ces noms peuvent être un *IsmFaîl* (*i.e.* participe actif), un *IsmMafôul* (*i.e.* participe passif) ou un *Masdar* (*i.e.* forme infinitive) (Dichy et Farghaly, 2003).

(3) Nous avons introduit dans le même dictionnaire quelques mots au pluriel qui n'ont pas de correspondant singulier utilisé ; par exemple, le mot ' مَخَاوف ' (*MakhAwif* – dangers, périls).

Des classes flexionnelles sont associées à l'ensemble des noms primitifs et aux déverbaux, afin de représenter l'ensemble des cas possibles (nominatif, accusatif et génitif) et l'ensemble des formes fléchies correspondantes (féminines, duales et plurielles). Notons ici que la déclinaison de l'ensemble des noms au pluriel a nécessité le développement de 65 modèles de flexionnels pour décrire les pluriels externes[9] formés par l'ajout d'un suffixe au singulier sans changement de la structure du mot et les pluriels internes[10] formés par modification de la structure interne du mot avec conservation des lettres de base.

5.3. Les particules

Nous avons répertorié environ 450 particules vocalisées. Ces particules incluent les prépositions, adverbes, conjonctions, interjections et les outils d'exceptions, de négation, etc.

La formalisation de la flexion des verbes, des noms primitifs et des déverbaux permet de reconnaître toutes les formes fléchies correspondantes ; l'algorithme de consultation de NooJ utilise des automates finis, ce qui permet de reconnaître directement leurs formes non voyellées et partiellement voyellées. De plus, chaque forme reconnue est associée par l'algorithme de consultation de NooJ à des informations linguistiques : lemme, catégorie grammaticale, genre, nombre, informations syntaxiques (*e.g.* +Transitif) et distributionnelles (*e.g.* +Humain).

6. Analyse morphologique et définition des règles grammaticales

La langue arabe étant une langue fortement agglutinante, son analyse morphologique se déroule en deux phases (cf. figure1).

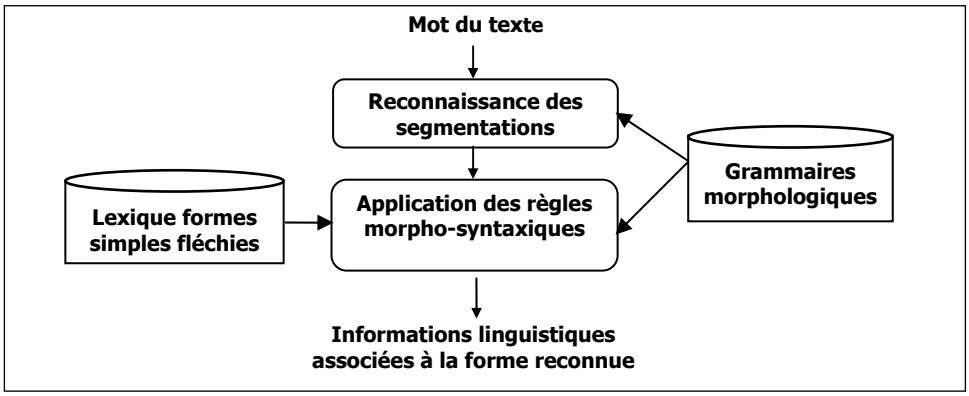

Figure 1. Chaîne d'analyse morphologique d'un mot

[9] Les pluriels externes sont les pluriels réguliers masculins et les pluriels réguliers féminins.
[10] Les pluriels internes sont les pluriels brisés et les pluriels quadrisyllabiques.

Dans un premier lieu, un système de décomposition des formes, implémenté sous forme de transducteurs finis (grammaires morphologiques NooJ), est appliqué à chaque forme du texte pour reconnaître les segmentations potentielles en identifiant le radical et les différents affixes qui lui sont collés. En second lieu, une phase d'application de règles morpho-syntaxiques associe la reconnaissance d'une forme à un ensemble de contraintes lexicales permettant de travailler uniquement avec des combinaisons valides des différents constituants de la forme. Les segmentations retenues sont validées grâce à une consultation du lexique des formes simples fléchies précédemment construit. À l'intérieur des grammaires morphologiques, nous avons implémenté quatre types de contraintes lexicales :

- **Contraintes morphologiques :** ces contraintes découlent de l'altération de certains radicaux par agglutination à un préfixe ou suffixe. Elles permettent de rétablir la graphie initiale, telle qu'elle figure dans le lexique. Ces contraintes tiennent compte des incompatibilités morphologiques qui auraient être générées, à partir d'une décomposition directe, par le biais de transformations morphologiques (ajout de lettres, suppression, substitution, etc.).
- **Contraintes sur les propriétés syntaxiques des verbes** : ces contraintes prennent en compte la marque « +Transitif » des verbes du dictionnaire. En effet, la transitivité des verbes nous permet, généralement, de décider l'agglutination des suffixes aux verbes. Une telle agglutination ne sera permise que pour les verbes transitifs directs ou transitifs indirects conjugués à la 3e personne du singulier (Achour, 1998)

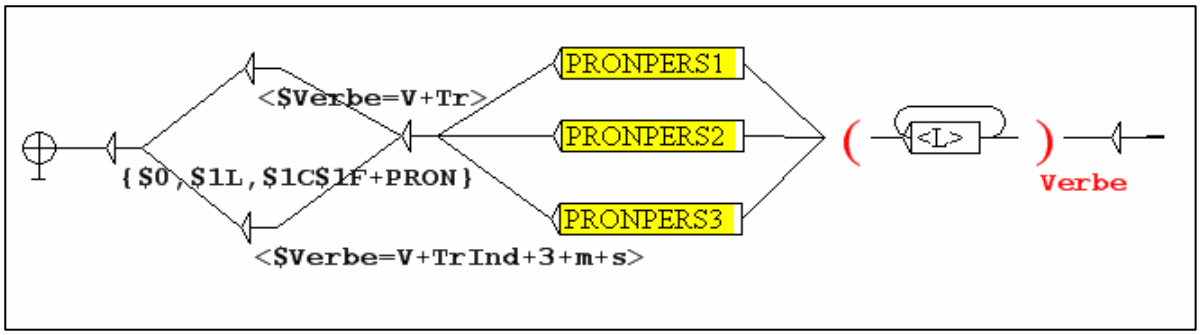

Figure 2. Contraintes lexicales : la transitivité des verbes

Le graphe simplifié ci-dessus montre que l'acceptation d'une entrée formée par l'agglutination d'une suite de lettres (<L>) sauvegardée dans la variable ($Verbe), suivie par un pronom personnel (PRONPERS1, PRONPERS2, PRONPERS3)[11] est liée à la vérification de l'une des deux contraintes lexicales ; d'une part la variable '$Verbe' peut représenter n'importe quelle forme fléchie d'un verbe transitif (<$Verbe=V+Tr>), d'autre part une forme conjuguée à la 3e personne (3) du masculin (m), singulier (s) d'un verbe transitif indirect (+TrInd) (<$Verbe=V+TrInd+3+m+s>).

- **Contraintes orthographiques :** ces contraintes prennent en compte le changement de l'orthographe de certaines lettres lors d'une agglutination. Nous citons le cas de la lettre ' t ' acceptant deux orthographes différentes ('ت','ة') et le cas des 'alifs' possédant cinq orthographes différentes ('ء','أ','إ','ؤ','ئ'). La présence de l'une ou l'autre des orthographes potentielles est relative à la nature de la forme et la position de la lettre dans le mot. Des opérations de substitution sont, alors, prévues avant la consultation du dictionnaire et l'association des informations linguistiques correspondantes à la forme en entrée.

[11] Les nœuds colorés (PRONPERS1, PRONPERS2, PRONPERS3) représentent des sous-graphes.

- **Contraintes phonologiques :** ces contraintes, généralement combinées avec les contraintes ci-dessus dans les grammaires morphologiques, permettent de maintenir une consonance harmonieuse à l'intérieur des formes agglutinées. Elles concernent la compatibilité de la flexion casuelle du radical avec celle du suffixe qui s'y rattache.

7. Conclusion : résultats et perspectives

Cet article décrit une formalisation du vocabulaire de l'arabe standard et une chaîne d'analyse morphologique de formes ; cette dernière est indépendante de l'état de vocalisation et d'agglutination de celles-ci. Le module construit est utilisé, au sein de l'environnement linguistique de développement NooJ, pour la restitution des voyelles manquantes et l'analyse linguistique des écrits arabes. L'évaluation de la couverture lexicale de ce module est entreprise en effectuant l'analyse lexicale du corpus du LASELDI récupéré à partir d'Internet. Ce corpus est composé d'articles journalistiques du journal *Le Monde Diplomatique*[12], qui comporte environ 150 000 formes différentes. Le résultat de l'analyse lexicale montre que le vocabulaire du corpus est reconnu à 93 % par nos ressources lexicales et morphologiques. L'ensemble des formes non reconnues contient 7 000 formes translittérées de noms propres (*e.g. Chirac, Marseille, UNICEF*, etc.) ainsi que quelques dérivations (*e.g. Chiraquisme*), deux milliers d'emprunts et environ 1 400 fautes d'orthographe. La majorité des formes non reconnues sont des noms propres de personnes, d'organisations ou de localités ; il nous faudra maintenant, d'une part, implémenter un module de reconnaissance de ces entités nommées et, d'autre part, valider les formes reconnues (nous utilisons pour cela des grammaires locales syntaxiques).

Références

ABOU IL AZM A. (2003). *Tasrif Moojim il afál : 10 000 verbes*. Dar Ittawhidi, Rabat.

ABDELI A., COWIE J., SOLIMAN H. (2004). « Arabic Information Retrieval Perspectives ». In *Actes de JEP-TALN, Analyse Automatique de l'arabe écrit et parlé*.

ACHOUR H. (1998). *Contribution à l'étude du problème de la voyellation automatique de l'arabe*. Thèse de doctorat, Université Paris 7.

BEESLEY K., BUCKWALTER T. (1989). « Two-level Finite State Analysis for Arabic Morphology ». In *Actes du séminaire On Bilingual Computing in Arabic and English.* Cambridge.

BEESLEY K. (1996). « Arabic Finite-State Morphological Analysis and Generation ». In *Actes de COLING96*. Copenhagen.

BEESLEY K. (1998). « Arabic Morphology Using Only Finite-State Operations ». In *Actes de Approches Informatiques pour le traitement des langues sémitiques*. Montréal.

BEESLEY K. (2001). « Arabic Finite-State Morphological Analysis and Generation of Arabic at Xerox Research : Status and Plans in 2001 ». In *Actes de ACL/EACL2001*. Toulouse.

DEBILI F. (2001). *Traitement automatique de l'arabe voyellé ou non.* Correspondances–IRMC.

DICHY J., FARGHALY A. (2003). « Roots & Patterns vs. Stems plus Grammar-Lexis Specifications: there what basis should is multilingual lexical database centred on Arabic be built? ». In *Journée On Machine Translation for Semitic Languages*. New Orleans.

KOULOUGHLI D.E. (1994). *Grammaire de l'arabe d'aujourd'hui*. Édition Perfectionnement

KOSKENNIEMI K. (1983). *Two-level morphology : a general computational model for word-form recognition and publication*. Publication N° 11. Université de Helsinki.

[12] Le corpus a été téléchargé, en majeure partie, à l'adresse : http://www.mondiploar.com

MCCARTHY J. (1981). « A prosodic theory of non concatenative morphology ». In *Linguistic Inquiry* 12 (3) : 373-418.

REVUZ D. (1991). *Dictionnaires et lexiques : méthodes et algorithmes*. Thèse de doctorat, Université Paris 7.

SILBEZTEIN M. (2005). « NooJ's Dictionnaries ». In *Actes de LTC 2005*. Poznan.

TUERLINCKX L. (2004). « La lemmatisation de l'arabe non classique ». In *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*. Presses universitaires de Louvain, Louvain-la-Neuve.