VERS UN SYSTEME D'ANALYSE MORPHO-SYNTAXIQUE DE LA LANGUE ARABE TOLERANT LES FAUTES

Nacéra TAIBI
Centre de recherche scientifique et technique
pour le développement de la langue arabe

Résumé

S'insérant dans le cadre du traitement automatique de la langue arabe, le présent article décrit un système d'analyse morpho-syntaxique tolérant les erreurs pour une interface d'interrogation de base de données. Cette analyse de l'arabe puise son formalisme des concepts du modèle linguistique néo-Khalilien. Ce dernier permet de concevoir une analyse morpho-lexicale dérivationnelle et une analyse syntaxique fondée sur des structures dites schèmes générateurs. Ce système est muni de deux correcteurs de fautes : l'un pour les erreurs lexicales et s'appuie sur une approche basée sur les combinaisons phonologiques interdites de l'arabe et une variante des techniques de renversement ; l'autre est utilisé pour certaines erreurs d'accord résolues grâce aux réseaux de transitions augmentés ATN, outil permettant la mise en œuvre du formalisme utilisé dans la syntaxe.

Mots-clés : Traitement automatique de l'arabe - analyse morpho-syntaxique - erreurs lexicales - erreurs d'accord - techniques de renversements - ATN - modèle linguistique néo-Khalilien.

الملخص

يدخل هذا البحث في إطار العلاج الآلي للغة العربية حيث يتعرض هذا المقال إلى وصف نظام تحليلي صرفي نحوي يصحح الأخطاء المستعملة لغرض تصميم واجهة استفهامية لقاعدة معطيات، ويستمد هذا التحليل صيغته الصورية من مفاهيم النموذج اللساني للنظرية الخليلية الحديثة التي تسمح بوضع نظام تحليل صرفي تفريعي وآخر تركيبي، ويرتكز على ما يُسمى بالحدود الإجرائية. وهذا النظام مزود بمصححين للأخطاء: المصحح الأول خاص بالأخطاء اللفظية، ويرتكز على طريقة تعتمد على التراكيب الفونولوجية غير المقبولة في اللغة العربية، بالإضافة إلى إحدى المتغيرات الخاصة بتقنيات القلب. أما المصحح الثاني فيُستعمل في أخطاء الربط التي يتم تحليلها بواسطة شبكات ATN التي تسمح بالبرمجة الآلية للصيغة الصورية المستعملة في مستوى التراكيب.

الكلمات المفاتيح: العلاج الآلي للغة العربية – تحليل صرفي نحوي– الأخطاء اللّفظية– أخطاء الرّبط – تقنيات القلب – ATN –النموذج الخليلي الحديث .

Abstract

As part of the framework of arabic automatic processing, this paper describes a system of a morpho-syntactic analysis that allows errors for a database interface. The formalism of this arabic language is based on concepts of the neo-khalilian linguistic model. This latter allows to conceive a derivational morpho-lexical analysis and syntactic analysis both based on structures known as generating schemes. This system contains two error correctors: one for the lexical errors, this one is based on an approach that uses unallowed phonological combinations of arabic and a variant of inversement technics; the other corrector is used to resolve some agreement errors thanks to the augmented transitions network ATN which is a tool to implement the formalism used in syntax.

Keywords : Arabic automatic processing - morpho-syntactic analysis - lexical errors - agreement errors - inversement technics – ATN – neo-khalilian linguistic model.

Introduction

Le traitement automatique des langues naturelles (TAL) est la discipline dont l'objet est la création de programmes informatiques capables d'interpréter ou de produire des phrases en langage naturel. Ce traitement est cependant d'une telle complexité qu'il est considéré comme un processus cognitif mettant en relation différentes disciplines telles l'intelligence artificielle, la linguistique, la psychologie et la philosophie. Le champ scientifique lui correspondant est *l'informatique linguistique*. L'interface avec l'ordinateur en langue naturelle, outil bien agréable pour une communication Homme/Machine est l'une des applications du TAL. S'inscrivant dans cet axe de recherche, le présent article décrit un travail ayant pour objectif l'élaboration d'un analyseur morpho-syntaxique de l'arabe qui se veut robuste et efficace, et son intégration dans une application d'interface d'interrogation d'une base de données textuelles. Cette interface étant amenée à être utilisée par un large éventail de personnes, le problème de la tolérance de certaines erreurs apparaissant dans les requêtes est posé ; notre système se charge alors d'améliorer la convivialité de cette dernière en essayant de faire face aux erreurs inhérentes à l'usage de la langue naturelle. Nous nous intéressons pour cela aux erreurs lexicales de type erreurs de frappe et d'orthographe, et aux erreurs de grammaire de type erreurs d'accord.

Le traitement automatique que nous faisons s'applique à l'arabe écrit non voyellé, tel qu'on le rencontre dans les documents. Le fait de ne pas introduire les voyelles constitue une difficulté supplémentaire qu'il faut prendre en considération.
Notre système d'analyse morpho-syntaxique [TAIBI 97] est basé sur le modèle linguistique néo-khalilien [HADJ-SALAH 79] dont l'exploitation confère au module syntaxique des fondements plus solides.

Nous décrivons dans cet article les principaux concepts de la théorie néo-khalilienne portant sur la syntaxe, puis la démarche adoptée pour aborder le problème de la correction d'erreurs lexicales qui va de pair avec l'analyse morpho-lexicale ; nous présentons ensuite la mise en œuvre de l'analyse syntaxique grâce à l'outil linguistique que sont les réseaux de transitions augmentés qui nous permettent de décrire les connaissances syntaxiques et de résoudre quelques problèmes d'erreurs d'accord.

I. Le modèle linguistique néo-khalilien

La théorie néo-khalilienne, développée par le professeur HADJ-SALAH [HADJ-SALAH 79], s'est inspirée des travaux des anciens grammairiens arabes, à savoir *Al-ẖalīl Ibn Aḥmad Al-Farāhīdī* et son disciple *Sībawayh*. Ces recherches avaient pour but d'aboutir à des fondements scientifiques pour une description de toute la langue et du naḥw (grammaire) en particulier.

Ce modèle linguistique est orienté vers la reconnaissance de structures syntaxiques explicitées par le biais de schèmes générateurs, il décrit pour cela la langue arabe comme étant constituée de trois niveaux, à savoir :

- le niveau inférieur correspondant à la kalima,
- le niveau intermédiaire, dit intra lexical, est celui de la lexie,

- le niveau supérieur représentant la tectonie en particulier et regroupant la syntaxe en général.

1. La kalima

La notion de kalima (pluriel kalims) intervient au niveau de la morphologie et est définie comme étant un nom, un verbe ou un mot outil. Ces kalims sont irréguliers tels les mots outils ou les noms propres et communs, ou réguliers et dans ce dernier cas ils obéissent à des règles de formation en racines et schèmes tels les verbes et certains noms.

1.1. Notion de racine

Soit l'exemple suivant : كَتَبَ، كَـاتِبٌ، مَكْتُـوبٌ : il y a des consonnes invariablement à la même position avec certains ajouts, à savoir (ك، ت، ب). Ces éléments communs correspondent à la racine définie comme étant une séquence de consonnes primitives non ajoutées qui forment un ensemble ordonné.

1.2. Notion de schème

Cette notion de schème n'est rencontrée que dans les écrits des anciens grammairiens arabes et des orientalistes contemporains, elle est donc propre à la linguistique arabe.

Al-Raḍi al-istrābādī met en évidence cette notion de schème dans la définition suivante : «On entend par *binā'*, *Wazn* ou *ṣīġa* (= respectivement : structure, module et forme), d'une *kalima* (lexème) la *configuration* (hay'a) qu'elle possède et que peuvent posséder en commun avec elle (yu-šārikuhā) d'autre lexèmes. Elle consiste dans :

- le *nombre* de ses segments consonantiques *ordonnés* (murattaba),
- les phonèmes vocaliques pleins et *bien déterminés* et les voyelles zéro (sukūn) [qui accompagnent ces segments], ceci compte tenu [des correspondances] entre segments adventices [d'une part] et radicaux [d'autre part] ; tout élément [étant considéré par ailleurs] *selon la position qu'il occupe*» [HADJ-SALAH 79].

Soit les exemples suivants : مَكْتَب مَرْكَب مَقْعَد. En considérant les consonnes de la racine comme variables et en les remplaçant par des C_i, i = 1, 2, 3, on obtient : $maC_1C_2C_3$ équivalent à مَفْعَل qui n'est que le schème de ces trois kalims. Le schème est donc une séquence virtuelle dans laquelle la racine s'intègre pour former une kalima.

2. La lexie

La lexie est toute séquence isolable et indivisible qui admet ou non des ajouts par simple concaténation sans que cela lui fasse perdre son caractère de séquence insécable du point de vue de sa réalisation. Elle est donc constituée d'un aṣl ou noyau omniprésent dans la structure syntaxique telكتـاب, pouvant accueillir d'une manière alternée et/ou continue des unités incrémentielles dans des positions bien définies tel كتاب العقاد في. Ces positions forment un ensemble structuré qui constitue ce qui est nommé SCHEME GENERATEUR.

Il existe deux types de lexies : nominale et verbale.

a- La lexie nominale est caractérisée par un schème générateur bi-dimensionnel : des incrémentations par antéposition ou postposition par rapport à un noyau sur l'axe horizontal sont possibles, de plus ces transformations sont réversibles sur l'axe vertical.

Ce schème générateur est illustré par les positions suivantes :
[particule du génitif] + [Identifiant] + **NOYAU** + [Désinences] + [Tanwīn] / [Complément adnominal] + [Caractérisant].
Exemples: {(noyau) عمـــر }; {(complément adnominal) الدار (noyau) صاحب}; {(noyau) بحر (identifiant) الـ (particule du génitif) في} {(caractérisant) شامخة (noyau) جبال}.
Les positions du complément adnominal et du caractérisant sont des zones de récursivité où des enchâssements de séquences sont possibles. Dans les exemples suivants, on remarque une récursivité linéaire : au niveau du caractérisants {[(caractérisant) ظريف شـــيق] (noyau) كتـــاب} ; au niveau du complément adnominal {[(complément adnominal) مدرســـــة الحيّ] (noyau) معلم}.
b- la lexie verbale est définie par trois schèmes générateurs différents qui correspondent aux trois modes verbaux l'accompli, l'inaccompli et l'impératif. Nous les regroupons dans cette syntaxe où certains ajouts peuvent ne pas exister en fonction du mode considéré.
[Exposant] + [Convertisseur] + **noyau verbal (fi'l + ḍamīr)** + [Marque flexionnelle] + [pronoms affixes].
Exemples : {(noyau) خرجـــت (exposant) قـــد}; {(pronom affixe) هـــا (noyau) يودعـــو (convertisseur) كي}; {(noyau) إقرأ}.

3. La tectonie

La tectonie syntaxique est définie par des séquences constituées de lexies. Cependant, le lien regroupant ces séquences n'est pas une simple concaténation car la suppression de l'une de ces lexies fait disparaître l'unité à ce niveau. Le schème générateur qui décrit une tectonie, plus abstrait que celui de la lexie, est représenté par un couple ordonné (Régissant, Termes régis) équivalent à la formule {R T_1 T_2}, où les éléments (R, T_1) sont plus liés que (R, T_2) ou (T_1, T_2).
Les T_i sont des lexies, et le régissant R correspond à :
-zéro ou ibtidā' T_2 = ربيعي T_1 = الطقس R = ø,
-un exposant non verbal (classe de إن): T_2 = سريع الحركة T_1=الفتى R = إن,
-un verbe exponentiel (classe de كان) : T_2 = رائعة T_1 = الطريق R = كانت,
-un verbe non exponentiel : T_2 = كنزا T_1 = السائح R = وجد.
Le niveau syntaxique ne se limite pas à une tectonie car à cette dernière peut être ajouté un déterminant D de différentes natures : D = متألما T_2 = صوته T_1 = الخدم R = رفع.
Une récursivité par enchâssement existe aussi à ce niveau et ce dans chacune des positions R, T_1, T2 et D.
Dans l'exemple suivant R est une tectonie : T_2 = مريضا T_1 = عليا R = [T_2 = المعلم T_1 = ت R = أعلم].
Dans cet autre exemple, T_2 est une tectonie : T_2 = [T_2 = ذاك T_1 = ø R = يقول] T_1 = عمر R = ø.
Il existe un autre niveau dit supralexical où des mots outils peuvent apparaître à l'initiale de séquences ayant la structure RTi D.

II. L'analyse morpho-lexicale

L'analyse morpho-lexicale [TAIBI 94] permet d'obtenir des informations linguistiques (catégorie, genre, nombre) associées aux entités lexicales dans une phrase. Cependant, la particularité de la langue arabe fait qu'une entité lexicale peut être constituée de plusieurs kalims tels:ليدرسونه بمدرستهم. Nous analysons l'entité dans ce cas selon une micro-syntaxe déduite du schème générateur des lexies, elle est alors supposée être formée de kalims antéposés et/ou postposés à un noyau.

L'entité peut être aussi une kalima dérivable, notre traitement ne se limite alors pas à vérifier son existence dans le lexique comme c'est le cas pour les mots outils ou les noms propres et communs, mais il s'appuie sur une analyse morphologique dérivationnelle qui se charge de retrouver le schème et la racine dont la kalima dérive. Pour ce type de kalims le lexique ne conserve que la forme jugée représentative d'une kalima dérivable, à savoir son schème et sa racine ; il est alors moins volumineux.

Pour une kalima, retrouver automatiquement la racine et le schème dont elle dérive revient à retrouver le schème de même longueur que la kalima ; puis à en extraire la racine en récupérant les consonnes de la kalima de même rang que les Ci ; la vérification de l'existence de la racine récupérée est ensuite réalisée grâce à une recherche dans le dictionnaire des racines classées par ordre alphabétique ; si la racine existe, nous vérifions la compatibilité entre le schème et la racine en identifiant la classe de ce schème dans la liste de ceux répertoriés avec la racine, évitant ainsi toute décomposition erronée.

Un traitement particulier est réservé aux racines possédant des voyelles longues, une hamza ou avec une gémination ; les notions de schème de surface et de schème profond sont introduites avec certaines règles de transformation pour permettre de retrouver la racine dont la kalima découle.

Exemples : - pour قال, son schème de surface est فال, il n'est pas attesté par la langue mais déduit de la kalima ayant subi des transformations. Ce dernier appartient à une classe de schèmes ayant pour schème profond فول qui nous permet d'extraire la racine قول.

- La kalima اتصل répond au schème de surface عل + ا ; une règle de transformation indique que dans ce cas + remplace une consonne de l'ensemble {ت ا و ي} correspondant à la première radicale dans le schème افتعل, la racine récupérée est وصل.

Prise en hors contexte, une kalima peut avoir plusieurs interprétations, c'est le cas de أكل :

- (أكل، فَعْل) : مصدر مذكر مفرد
- (أكل، فَعَل) : فعل ماضي، مبني للمعلوم، مجرد مصرف في المفرد المذكر لضمير الغائب
- (أكل، فُعِل) : فعل ماضي، مبني للمجهول، مجرد مصرف في المفرد المذكر لضمير الغائب

Cette ambiguïté lexicale est levée selon le contexte au niveau des analyses syntaxiques et sémantiques.

Par contre pour الأكل, l'analyse donnée : الـ : أداة التعريف

(أكل، فعل) مصدر مذكر مفرد

Nous n'avons retenu pour أكل que l'interprétation de nom car nous avons défini des règles contextuelles qui assurent la compatibilité entre les différentes interprétations de kalims

formant une entité lexicale, dans ce cas la règle dit qu'après ال seule l'interprétation de nom est retenue.

4. Lexique

Le lexique est constitué de trois dictionnaires :

- le dictionnaire des kalims non dérivables (les mots outils et les noms propres et communs) ;
- le dictionnaire des racines: à une racine correspond sa chaîne consonantique et les classes des schèmes qu'elle accepte ;
- le dictionnaire des schèmes : à un schème correspond sa classe de schème profond et un ensemble de diacritisations, à chacune d'elles sont associés sa classe et un ensemble de traits syntaxico-sémantiques.

Les dictionnaires des kalims non dérivables et des racines utilisent un accès indexé sur leur ordre alphabétique. Le dictionnaire des schèmes est organisé en fichiers, où chacun d'eux contient un ensemble de schèmes de même taille car leur traitement nécessite un accès séquentiel par longueur.

III. La stratégie de correction d'erreurs lexicales

Une entité lexicale qui n'a pu être analysée est à l'origine du processus de correction. Nous nous proposons de corriger une seule erreur dans l'entité et nous nous limitons aux erreurs d'orthographe et de frappe, type qu'on ne peut d'ailleurs pas dissocier. Ces fautes peuvent être causées par l'omission ou l'insertion d'un caractère, la substitution d'un caractère par un autre ou la permutation de deux caractères consécutifs.

Pour traiter ces erreurs lexicales nous avons mis en œuvre une stratégie de correction qui se base sur une technique de détection et sur une variante des méthodes de renversement tout en prenant en compte la morphologie dérivationnelle d'une entité lexicale.

1. La technique de détection

Cette technique est basée sur les combinaisons phonologiques interdites de certains phonèmes arabes et permet de localiser l'erreur dans l'entité. Elle s'appuie sur une classification conçue à partir des travaux des anciens grammairiens arabes sur les combinaisons de phonèmes acceptées ou non par la langue arabe.

L'approche de détection est constituée d'un ensemble de règles de différents types. Par exemple, soit la graphie كلأ à corriger, nous lui appliquons le type de règles qui permet de retrouver l'incompatibilité entre deux consonnes successives dans une entité. La règle adéquate est la suivante : si A = {ء ئ ؤ} alors $\forall x \in A$, le successeur de $x \notin A$. l'erreur est alors localisée au niveau du segment لأ et l'un de ces deux caractères est en plus ou bien il manque un caractère entre les deux.

2. La technique de correction

Cette technique est basée sur le renversement d'erreurs [POLLOCK 84] [MARET 87]. Elle s'appuie sur l'idée qu'une chaîne erronée peut être transformée en un mot appartenant au lexique par l'inversion d'une des opérations d'édition (insertion, omission, substitution, transposition), alors ce mot est une correction possible pour la chaîne en question. Notre traitement consiste à :

- appliquer l'algorithme de détection pour éventuellement localiser l'erreur ;
- générer l'ensemble des chaînes voisines de la chaîne erronée à une erreur près, pour une omission ou une substitution, le caractère * remplace le caractère manquant ou erroné ;
- retrouver ensuite parmi les chaînes de cet ensemble celles qui existent dans le lexique et qui sont des solutions candidates.

Exemples : عــأكل et قــظ. Nous appliquons d'abord la procédure de détection à la graphie erronée. Dans le cas où l'erreur est localisée – avantage de l'outil de détection – nous ne générons pour la graphie que les chaînes voisines à une erreur près en cette position. Pour عأكل nous nous limitons aux chaînes de correction possibles suivantes : insertion (عكل،أكل), omission (ع*كل), substitution (أكل* ،ع*كل), transposition (عأكل).

Dans le cas où l'erreur n'est pas localisée nous générons pour la graphie erronée l'ensemble des chaînes qui lui sont voisines à une erreur près. Pour l'exemple قظ nous avons : omission (ق*ظ ، قظ ، قظ*) substitution (ق*، ظ*), transposition (ظق).

Ensuite nous cherchons les solutions candidates en comparant ces chaînes de correction possibles aux entrées du lexique. Nous nous intéressons d'abord aux dictionnaires des kalims non dérivables, pour قــظ les solutions proposées sont {قظ، قد}. Pour les chaînes de correction possibles non sélectionnées par le précédent traitement nous vérifions leur dérivabilité en schème et racine, ce qui donne pour عأكل les solutions { أكل، يأكل، تأكل، نأكل}.

Lorsque le système n'offre aucune solution ou que l'utilisateur ne retient pas les solutions précédemment proposées, la chaîne erronée est considérée comme constituée de plusieurs kalims, l'erreur peut être située au niveau de l'un des kalims. L'algorithme de correction arrive à isoler le segment affecté et le corrige à part, ce qui permet de faciliter et d'optimiser le processus de correction des erreurs lexicales. Pour l'exemple عأكل en recherchant l'erreur dans le segment antéposé et postposé au noyau, les solutions candidates sont {فأكل، بأكل لأكل، أأكل }.

Cette technique est intéressante car elle permet un accès tolérant dans un lexique de grande taille : les accès lexicaux sont restreints à l'ensemble des chaînes générées, la correction de la chaîne erronée est équivalente à l'intersection de cet ensemble et du lexique. Elle a aussi pour avantage de corriger aussi bien les mots longs que les mots courts, de plus elle n'est pas sensible aux erreurs portant sur une position donnée dans un mot.

IV. L'analyse syntaxique

1. Formalisme

L'analyse syntaxique est un passage obligé pour réaliser des systèmes élaborés de traitement automatique des langues [FAY-VARNIER 91]. Notre approche syntaxique est fondée sur le modèle linguistique néo-khalilien qui fournit une description générale et cohérente de la langue arabe, entre autres pour les structures syntaxiques enchâssées. Son utilisation évite de se baser sur la grammaire traditionnelle et donc sur un mécanisme ad hoc. Ce formalisme est mis en œuvre avec un modèle calculatoire, nous avons opté pour les réseaux de transition augmentée, technique qui s'est révélée très efficace notamment pour les interfaces de bases de données en langue naturelle où les résultats sont probants [SABAH 89]. Cet outil est à la fois une reformulation et une extension des grammaires à contexte libre. Dans la reformulation les règles de production sont traduites en des graphes qui mettent en facteurs des sous-graphes. L'extension correspond à l'utilisation des ATN au niveau fonctionnel, ils permettent de déterminer les relations entre les éléments porteurs de sens. On associe alors pour chaque graphe une description fonctionnelle, qui sert à mémoriser diverses informations complémentaires (le genre, le nombre, le sujet ou l'objet) sous la forme de registres constitués d'attributs et de valeurs. Ces derniers sont mis à jour grâce à des conditions et des actions qui peuvent être associées à chaque arc du graphe et qui ont pour rôle de construire la représentation de l'énoncé traité.

2. Analyse

L'analyseur suit un algorithme descendant, il consiste à lire les mots un par un de la droite vers la gauche et à essayer de joindre l'état initial à un état final du graphe principal, en passant par les graphes appelés. Un arc est franchi si le mot courant et/ou les mots précédents remplissent toutes les conditions portées par cet arc, ses actions éventuelles sont exécutées mettant à jour les attributs du graphe en cours. L'analyseur est non déterministe et permet ainsi de gérer la grande ambiguïté syntaxique de la langue ; il prend en compte les phénomènes de retours arrières et le fait de se retrouver devant une impasse.
L'analyseur produit un arbre syntaxique où à chaque appel d'un graphe un nœud est défini, considéré comme élément de base de la structure engendrée, il contient comme étiquette la catégorie syntaxique du graphe en question, et il regroupe tous les registres utilisés. L'algorithme de l'analyse est entièrement récursif.

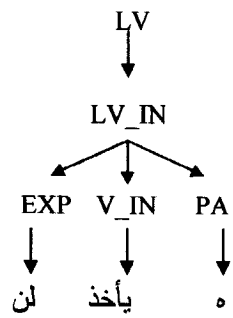
3. Grammaire

La grammaire que nous avons retenue est un sous-ensemble de l'arabe et a une couverture linguistique qui se limite à des phrases déclaratives, des questions et des requêtes pouvant contenir certains enchâssements. Elle est fondée sur l'exploitation des schèmes générateurs pour la reconnaissance des constituants syntaxiques d'une phrase. Elle est donc décrite en se basant sur les concepts de régissant (R), termes régis (Ti) et déterminant (D) et fait appel à des structures sous forme de lexies verbales et nominales.

3.1 Lexie verbale (LV)

Le graphe de la lexie verbale fait appel à trois sous graphes représentant le verbe à l'accompli, l'inaccompli et à l'impératif. Découlant du résultat de l'analyse morpho-lexicale d'une entité avec un ou plusieurs kalims post-posés, le verbe est suivi selon sa valence d'au plus deux pronoms-objet. Dans le cas de l'accompli et l'inaccompli, un ensemble de mots outils sont antéposés au verbe ; ils sont classés comme exposants ou convertisseurs. Pour l'inaccompli par exemple nous avons les convertisseurs (أن إذن كي) et les exposants (ل لما لم لن مالا س سوف قد).

Exemple : لن يـــأخذه Un exemple d'action dans le réseau est celle portée par l'arc PA et qui affecte le mot lu (ـــه) à l'attribut Objet.



3.2 Lexie nominale (LN)

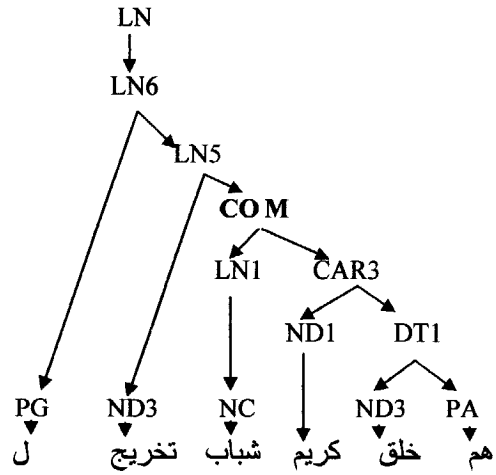
La grammaire de la lexie nominale est établie en fonction des valeurs de son noyau d'abord et en relation avec les structures syntaxiques de T_1 et T_2 ensuite.

Nous avons retenu vingt cinq valeurs pour le noyau, les plus usitées sont : pronom tonique, pronom affixe, démonstratif, adverbe, nom propre, commun, dérivé, de nombre et d'interrogation. Les graphes sont définis selon différents critères, à savoir :

- rassembler les types de kalims qu'on ne rencontre que dans la position du noyau comme le cas du pronom tonique.
- considérer l'indétermination et la détermination de la lexie (une lexie peut porter la détermination comme c'est le cas du nom propre, elle peut être déterminée par l'identifiant ال ou bien par le complément adnominal) ;
- selon que la lexie accepte ou non la classe des prépositions du génitif.

Les positions du complément adnominal ou du caractérisant, vu leur caractéristique de positions récursives ont fait l'objet de sous graphes à part.

Par exemple pour la position du complément adnominal, à l'exception du cas où le complément est un pronom affixe, cette position est susceptible de contenir des items répétés. Soit l'illustration suivante : لتخريج شباب كريم خلقهم. Le complément adnominal est une lexie nominale avec un déterminant خلقهم.



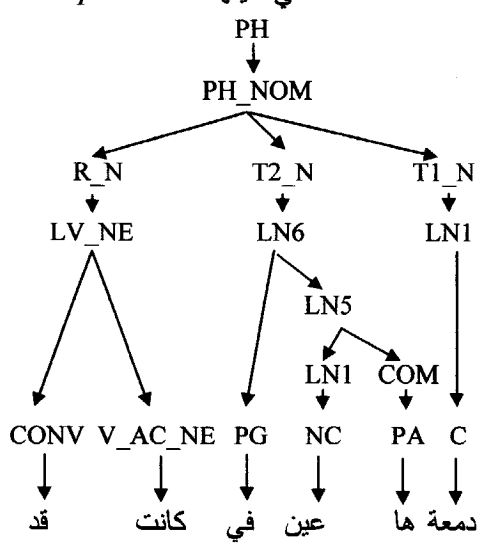
3.3. Niveau inter-lexical (PH)

Cette grammaire accepte deux types de phrases nominales et verbales: ph →ph_nom/ ph_verb structurés en RTi. L'ordonnancement des constructions de ces phrases dépend des valeurs du régissant et des termes régis, nous avons ainsi RT_1T_2, RT_2T_1 avec la condition que R soit différent de la classe de إن, T_2RT_1, RT_2 ou bien RT_1 dans le cas où le verbe est intransitif.

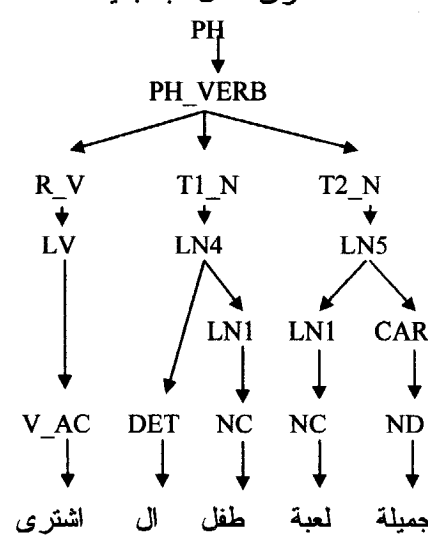
Dans une phrase nominale, le régissant R peut avoir pour valeur Zéro, un exposant non verbal ou un verbe non exponentiel. Les termes régis répondent à des structures particulières de lexies nominales et verbales. Par exemple T_1 n'est jamais un adverbe, ou bien T1 peut être un type particulier de lexie verbale <convertisseur = أن> + <verbe à l'inaccompli> ou bien T_1 ne contient jamais de particule de génitif.

Dans une phrase verbale, le régissant est une lexie verbale avec un noyau constitué d'un verbe intransitif ou transitif avec arguments. Les termes régis sont des lexies nominales.

Exemples : قد كانت في عينها دمعة

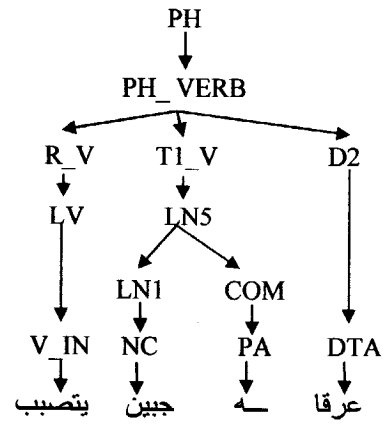


اشترى الطفل لعبة جميلة



Ces phrases nominales et verbales peuvent aussi contenir une classe D_1 déterminants comme le tamyīz, les phrases verbales acceptent une autre classe D_2 de déterminants, cas du ḥāl.

Exemple : يتصبب جبينه عرقا



3.3.1. La structure syntaxique interrogative

Une construction interrogative en arabe a le même ordre qu'une construction affirmative. L'interrogation est exprimée soit par :

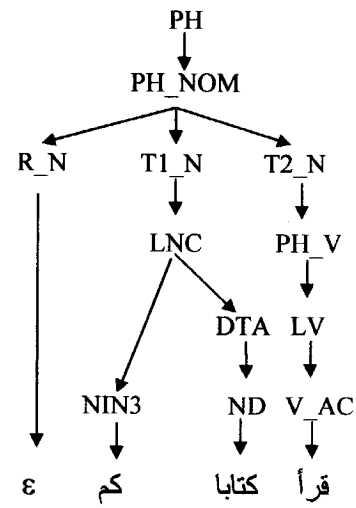
- les particules d'interrogation أ et هل qui sont des unités que l'on retrouve au niveau supra lexical et qui apparaissent à l'initiale de structures RTiD ;
- un nom d'interrogation, un adverbe ou un relatif qui sont des valeurs que l'on rencontre dans le noyau d'une lexie nominale, et qui sont en position de T_1 sujet.

Pour le nom d'interrogation كـــم, il introduit ce que nous avons appelé un déterminant de lexie qui est le tamyīz.

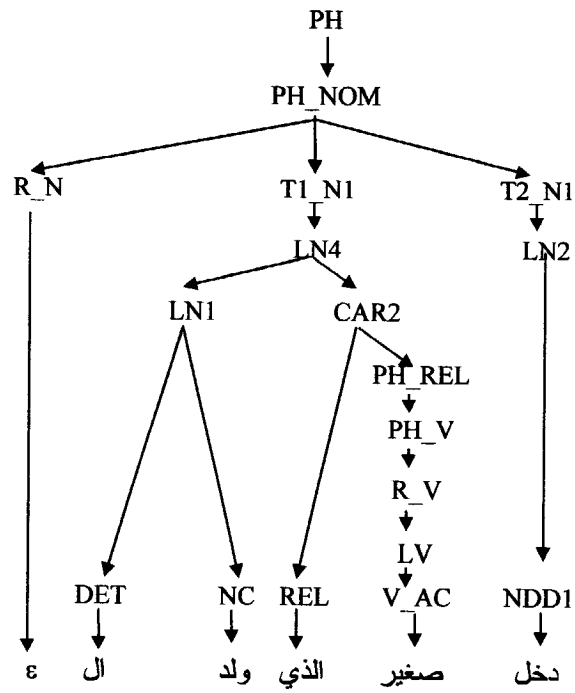
Soit l'exemple : كم كتابا قرأ؟

T_1_N est constitué d'une lexie nominale complexe formée d'un noyau كم et d'un déterminant de lexie كتابا.

Ainsi [كم كتابا] est le sujet et قرأ est l'attribut.



En plus de la récursivité linéaire, nous avons traité dans ce travail la récursivité par enchâssement au niveau de la tectonie donc dans les positions de R, T_1 et T_2 ; nous nous sommes limités à un emboîtement d'une lexie nominale dans la position du complément adnominal et nous avons pris en compte les relatives introduites par la classe de الذي.
Nous traitons ce type de constructions dans le cas où elles sont rencontrées dans des positions récursives au niveau de la lexie nominale à savoir celle du complément adnominal ou du caractérisant. Cette classe de الذي peut introduire une tectonie verbale ou nominale.
Exemple : الولد الذي دخـــل صغـير [الذي دخل] joue ici le rôle de caractérisant du noyau الولد d'une lexie nominale en position de T_1-sujet.



4. Erreur d'accord

Nous nous intéressons à trois cas d'erreurs d'accord dans une construction syntaxique:

- accord en genre et en nombre entre le sujet (T_1) et le verbe (R) et entre le nom (noyau) et son adjectif (caractérisant) ;
- accord au niveau de la rection du verbe à l'inaccompli par des mots outils (les convertisseurs ou exposants) qui se concrétisent par l'absence de marques flexionnelles à certaines personnes.

L'analyse de détection / correction des erreurs d'accord repose sur un certain nombre de principes décrivant les rapports entre les différentes composantes en termes de transmission de valeurs; nous avons regroupé les relations d'accord existant entre les différentes catégories sous formes de conditions et d'actions, puis nous analysons le fonctionnement de l'accord en termes de transmissions de valeurs. Nous avons distingué,

d'un point de vue fonctionnel, deux types d'accord, le premier entre deux catégories lexicales (convertisseurs ou exposant et verbe à l'inaccompli) et le second entre des catégories intra lexicales ou inter lexicales selon que nous sommes dans une lexie (noyau-nom et caractérisant-adjectif) ou dans une tectonie (régissant-verbe et terme régi 1-sujet). Par exemple pour contrôler l'accord en genre et en nombre entre le nom et son adjectif nous devons vérifier la contrainte d'accord, pour ce faire nous avons associé aux graphes de la lexie nominale des registres portant sur la détermination, le genre et le nombre . Les règles d'accord sont écrites sous forme de conditions et actions qui sont rattachées aux arcs définissant le nom et l'adjectif. Dans l'exemple suivant : الكتـــــاب المفيـــدة, une erreur d'accord en genre entre le noyau et son caractérisant est détectée, le noyau est au masculin alors que le caractérisant est au féminin. Le système suggère de remplacer المـــفيـــدة par un qualifiant masculin car il prend comme première hypothèse le fait que le genre et le nombre du noyau sont dominants. Si l'utilisateur n'est pas satisfait par cette correction, on suppose alors que l'erreur est au niveau du noyau.

7. Conclusion

Un système-prototype de l'analyseur morpho-syntaxique tolérant les fautes est écrit en PCScheme, un dialecte de Lisp. Ce système ainsi conçu offre la possibilité de corriger certaines erreurs lexicales et fautes d'accord. L'approche que nous proposons permet de prendre en compte le caractère dérivationnel de la langue arabe ainsi que ses particularités phonologiques dans l'analyseur morpho-lexical avec correcteur d'erreurs lexicales. De même, elle utilise dans la mise en œuvre de l'analyseur syntaxique les concepts linguistiques de la théorie néo-khalilienne qui définissent les composantes syntaxiques d'une phrase selon la notion de schèmes générateurs.

Le fait que nous utilisions un analyseur morpho-lexical complet et déconnecté du système global nous permet d'envisager sa portabilité vers d'autres applications ou analyseurs syntaxiques de la langue arabe. Une première application mise en pratique est un produit d'enseignement de la conjugaison et du système dérivationnel de la langue arabe.

Références

[COURTIN 90] COURTIN J., DUJARDIN D., KOWARSKI I., GENTHIAL D., STRUBE DE LIMA V. *Vers un système complet de détection / correction d'erreurs.* Rapport de recherche de l'équipe TRILAN, LGI-IMAG, Grenoble, Mai 1990.

[FAY-VARNIER 91] FAY-VARNIER C, FOUQUERE C., PRIGENT G., ZWEIGENBAUM P., *Modules syntaxiques des systèmes d'analyse du français.* TSI, Vol 10, n°6, 1991.

[HADJ-SALAH 79] HADJ-SALAH A. *Linguistique arabe et linguistique générale, Essai de méthodologie et d'épisthomologie du ilm al-Arabiyya.* 2 volumes, 1979.

[MARET 87] MARET D. *Comparaisons de chaîne de caractères, accès lexicaux tolérants et application.* Thèse de Doctorat, Université des Sciences Sociales, Grenoble, Mai 1987.

[POLLOCK 84] POLLOCK J., ZAMORA A. *Automatic spelling correction in scientific and scholary text.* CACM, Vol 27, n°4, Avr. 1984.

[PUJO 90] PUJO P., LANGELET G. *Aide intelligente pour la correction des requêtes en langage naturel à une base de données.* TSI, Vol. 9, n°3, 1990

[RICHARD 86] RICHARD D., LAPALME G. *Un système de correction automatique des accords des participes passés.* TSI, Vol 5, n°4, 1986.

[SABAH 89] SABAH G. *L'intelligence artificielle et le langage. Vol 2 : processus de compréhension.* Hermés, paris 1989.

[TAIBI 94] TAIBI N. *Automatisation de la dérivation lexicale arabe.* Rapport de recherche interne, CRSTDLA, 1994.

[TAIBI 97] TAIBI N. *Contribution à l'étude du traitement automatique des erreurs dans un texte écrit en arabe.* Thèse de Magistère, Ecole Normale Supérieure des Lettres et des Sciences humaines, 1997.

[WINOGRAD 83] WINOGRAD T. *Language as a cognitive process, voll : syntaxe.* Edition Addison Wesley, 1983.