# Quels attributs discriminants pour une analyse syntaxique par classification de textes en langue arabe ?

Fériel Ben Fraj (1), Chiraz Ben Othmane Zribi (2), Mohamed Ben Ahmed (3)
Laboratoire RIADI – Université La Manouba
ENSI, La Manouba, Tunisie
(1) Feriel.BenFraj@riadi.rnu.tn, (2) Chiraz.BenOthmane@riadi.rnu.tn,
(3)Mohamed.BenAhmed@riadi.rnu.tn

**Résumé** Dans le cadre d'une approche déterministe et incrémentale d'analyse syntaxique par classification de textes en langue arabe, nous avons prévu de prendre en considération un ensemble varié d'attributs discriminants afin de mieux assister la procédure de classification dans ses prises de décisions à travers les différentes étapes d'analyse. Ainsi, en plus des attributs morpho-syntaxiques du mot en cours d'analyse et des informations contextuelles des mots l'avoisinant, nous avons ajouté des informations compositionnelles extraites du fragment de l'arbre syntaxique déjà construit lors de l'étape précédente de l'analyse en cours. Ce papier présente notre approche d'analyse syntaxique par classification et vise l'exposition d'une justification expérimentale de l'apport de chaque type d'attributs discriminants et spécialement ceux compositionnels dans ladite analyse syntaxique.

**Abstract** For parsing Arabic texts in a deterministic and incremental classification approach, we suggest that varying discriminative attributes is helpful in disambiguation between structures to classify. That's why; we consider morpho-syntactic information of the current analyzed word and its surrounding context. In addition, we add a new information type: the compositional one. It consists of the portion of the syntactic tree already constructed until the previous analysis step. In this paper, we expose our parsing approach with classification basis and we justify the utility of the different discriminative attributes and especially the compositional ones.

**Mots-clés :** analyse syntaxique incrémentale, langue arabe, apprentissage automatique, classification, attributs discriminants

**Keywords:** incremental parsing, Arabic language, machine learning, classification, discriminative attributes

## 1 Introduction

Par intuition, l'analyse syntaxique d'une phrase par un humain ne correspond pas à la recherche de l'ensemble des règles nécessaires à la traiter mais plutôt à déceler toutes les connaissances disponibles au moment de l'analyse pour prendre la bonne solution ou celle qui semble être la meilleure. Nous suggérons, alors, qu'une analyse syntaxique automatique doit procéder de la même manière afin de faire correspondre les pré-requis théoriques (règles

grammaticales) et l'état du cas actuel à traiter pour aboutir à une bonne performance d'analyse.

Pour notre part, nous avons proposé pour la langue arabe (Ben Fraj et al., 2008a) une approche d'analyse syntaxique par classification à base d'apprentissage supervisé. Néanmoins, la prise de décision dans un modèle d'apprentissage automatique doit être conditionnée par un ensemble d'attributs dits d'apprentissage permettant la discrimination entre les différentes solutions (ou décisions à prendre) possibles. Pour ce faire, nous avons préconisé un ensemble varié d'attributs ; à savoir morpho-syntaxiques, contextuels et compositionnels. Mais, la question qui doit être posée à ce stade est : A quel point cette diversification peut-elle être utile et efficace pour notre procédure d'analyse syntaxique?

Le but de cet article consiste, alors, à valider expérimentalement l'apport des différents types d'attributs dans la procédure d'analyse syntaxique. Ainsi, nous commençons ce papier par la présentation de notre approche classificatoire d'analyse syntaxique de textes en langue arabe. Dans une seconde section, nous décrivons les différents attributs discriminants mis en jeu en apprentissage. La dernière section est consacrée à l'évaluation de la pertinence de ces attributs dans la tâche en main.

# 2 Analyse syntaxique de l'arabe comme un problème de classification

Notre approche d'analyse syntaxique se base sur l'attribution d'un ensemble d'arbres élémentaires aux différents mots d'une phrase par une procédure de classification, et ce en se basant sur leurs contextes respectifs. Ces arbres classes sont conçus suivant le formalisme grammatical d'arbres adjoints (TAG : Tree Adjoining Grammar). D'où la grammaire ArabTAG.

## 2.1 Présentation du formalisme TAG

La grammaire TAG (Joshi, 1987) fait partie des grammaires dites d'unification. Elle utilise la notion de structures de traits et peut être qualifiée de grammaire binaire vu qu'elle utilise deux structures arborescentes dites *arbres élémentaires* pour la représentation des éléments langagiers (mots ou catégories grammaticales) et deux opérations pour les accoler dans le but de générer deux arbres résultats.

Deux types d'arbres élémentaires sont utilisés:

- les arbres initiaux qui s'accolent à d'autres arbres par substitution
- et les arbres auxiliaires qui se lient avec d'autres arbres élémentaires ou fragments d'arbres dérivés par adjonction. Pour un arbre auxiliaire, le nœud racine doit avoir la même catégorie qu'un des nœuds feuilles du même arbre.

La grammaire TAG est qualifiée de simplicité vu que seules deux opérations de liaison entre structures sont permises :

- la substitution qui insère un arbre initial (ou le dérivé d'un arbre initial) à un nœud feuille d'un arbre élémentaire ou dérivé

- et l'adjonction qui permet la composition d'un arbre auxiliaire (ou dérivé d'un arbre auxiliaire) avec un arbre élémentaire ou dérivé. L'arbre auxiliaire, ayant deux nœuds particuliers racine et pied de même catégorie, peut s'adjoindre à un autre nœud racine ou interne de même catégorie.

A part l'arbre dérivé qui constitue l'arbre syntaxique de la phrase en cours de traitement, le formalisme TAG associe à toute phrase un arbre de dérivation où chaque nœud renvoie à une unité lexicale de la phrase et à la structure élémentaire choisie pour la représenter et chaque branche indique le type de combinaison entre les deux structures élémentaires qu'elle relie.

Ce formalisme grammatical d'unification a été accommodé (Ben Fraj et al., 2009) et utilisé pour la création de la grammaire ArabTAG.

## 2.2 La grammaire ArabTAG

ArabTAG (Ben Fraj et al., 2008b) est un modèle générique pour la représentation des structures syntaxiques modulaires de la langue arabe à base du formalisme précédemment présenté. Ces structures sont construites à partir d'une collection de règles grammaticales enrichies. L'enrichissement vient du fait qu'on a associé aux règles simples[1] d'une grammaire hors contexte certaines informations syntaxiques et sémantiques (se présentant les rôles des différentes composantes de la phrase) afin d'aider leur transformation en un modèle grammatical d'unification. Ces règles enrichies sont par la suite codées avec le langage de standardisation XML.

Ainsi, ArabTAG couvre la plupart des structures arborescentes standards utilisées dans les textes arabes. Lesdites structures sont organisées par familles suivant leur représentation d'une phrase (nominale (PN) ou verbale (PV)), d'un syntagme nominal (SN), d'un syntagme verbal (SV) ou d'un syntagme prépositionnel (SP). A l'intérieur de chaque famille, un ensemble assez varié et riche de sous-arbres est décrit. Pour la famille des SNs par exemple, nous avons conçu ses différents types ainsi que leurs représentations structurales possibles. Nous citons, à titre indicatif, le SN adjectival, le SN d'annexion, le SN quasi-propositionnel, etc.

## 2.3 Principe d'analyse

Notre analyseur syntaxique proposé est incrémental, déterministe et à base de classification. L'hypothèse incrémentale suppose que le traitement des langues humaines se fait dans une même direction : de droite à gauche dans notre cas, et procède par des découpages partiels qui permettent d'étendre les fragments composant une même phrase. Dans la stratégie incrémentale, le modèle d'analyse suit une chaîne de règles de prédictions autorisant le lien d'un nouveau mot avec la partie déjà construite de l'arbre d'analyse (Costa et al., 2001).

Considérons le mot $m_t$ en cours d'analyse à un instant $t$ du processus d'analyse et l'arbre partiel d'analyse $AP_{t-1}$ construit jusqu'à l'instant $t-1$. On cherche à trouver pour le mot $m_t$ l'arbre élémentaire convenable en se basant sur l'environnement du mot composé : des informations lexicales, syntaxiques et contextuelles relatives à ce mot et sur le fragment $AP_{t-1}$. D'autant plus que, nous présupposons que l'affectation des arbres élémentaires aux différents

[1] Les règles simples sont celles qui représentent seulement les composantes principales des phrases ; à savoir le sujet, le verbe et les compléments pour une phrase verbale et le sujet et le prédicat d'une phrase nominale.

mots de la phrase est intimement liée à la spécification des opérations nécessaires à les joindre ; c'est-à-dire que le choix d'un arbre élémentaire parmi d'autres pour un mot donné doit profiter à la fois du contexte lexical et de celui dérivationnel (ou compositionnel) des arbres élémentaires. Le résultat sera $AP_t$ qui est la portion d'arbre $AP_{t-1}$ incrémenté de l'arbre élémentaire prédit du mot $m_t$ à l'instant $t$.

# 3 Attributs discriminants

Les attributs discriminants constituent l'ensemble des descriptions qui permettent de différencier entre multiples cas à analyser. Ce sont les attributs susceptibles d'être pertinents pour le problème considéré. C'est ce qu'on appelle : langage de description du problème de classification. La question du choix des attributs est abordée lorsqu'il faut extraire des connaissances à partir de données ou de textes (Denis, Gilleron, 2000). Classifier un cas revient à chercher la probabilité de son appartenance à une classe sachant la description qui lui correspond. Dans la procédure de classification en main, les attributs pris en compte sont :

## 3.1 Attributs morpho-syntaxiques

Les attributs morpho-syntaxiques du mot en cours d'analyse qui sont variés : forme du mot[2] , valeur grammaticale (VG) positionnelle[3], lemme, valeurs et VGs des enclinomènes[4] s'ils existent, genre/nombre pour les noms et les adjectifs, temps/personne/transitivité pour les verbes (Ben Othmane, 1998). Ces informations permettent une bonne désambiguïsation entre les différents cas. Dans l'exemple de la figure 1, le fait que le mot يَقِلُّ est un verbe et qu'il est transitif direct et indirect à la fois permettent de choisir la structure d'une phrase verbale à deux compléments d'objet le premier est direct et le second est indirect.

## 3.2 Attributs contextuels

Les attributs contextuels qui rassemblent les caractéristiques morpho-syntaxiques des mots se trouvant au voisinage du mot en cours d'analyse. On peut utiliser une fenêtre (-k, +k) qui prend en considération le contexte à gauche et celui à droite. La valeur de k est définie par expérimentation. Ce genre d'informations permet de prendre en considération à la fois le passé et le futur de l'analyse. Ceci permettra d'améliorer la prédiction, d'éviter au maximum les erreurs d'analyse et aussi d'éliminer le risque des retours en arrière en cas d'erreur. Lors de l'analyse du mot cible دَوْرُ de l'exemple de la figure 1, la vérification du vocable se trouvant à la première position de son contexte à gauche et ses caractéristiques (proclitique et cas spécialement), permet de décider que les deux mots associés forment un SN d'annexion représenté par l'arbre❸.

[2] C'est la forme comme elle est présente dans le texte comme le mot الذهب (l'or).

[3] Exemples de valeurs grammaticales positionnelles (Ben Othmane, 1998) : اسم علم ـٌ (nom propre au nominatif) ou فعل مضارع معلوم ـَ (verbe au présent actif précédé d'un outil modificateur).

[4] Les enclinomènes agglutinés aux formes simples, permettent d'avoir des formes plus complexes comme dans l'exemple de la forme agglutinée : أستتذكروننا (est-ce que vous vous souviendrez de nous ?)

.

## 3.3 Attributs compositionnels

Ce sont les attributs compositionnels ou dérivationnels qui incluent la structure du fragment déjà construit de l'arbre d'analyse syntaxique à une étape intermédiaire de l'analyse. Ils englobent alors les nœuds (arbres élémentaires) déjà trouvés et spécialement ceux capables d'être dérivés dans des phases ultérieures d'analyse, les opérations possibles pour les lier et surtout les positions de jointure des anciens nœuds sur la portion construite de l'arbre d'analyse. Au fait, une décomposition en syntagmes est effectuée en cours d'analyse et ce sont les anciens points de jointure qui préciseront à quel ancien syntagme doit être accolé le mot en cours d'analyse ou doit-il plutôt constituer un nouveau syntagme. La figure 1 présente aussi cet aspect.

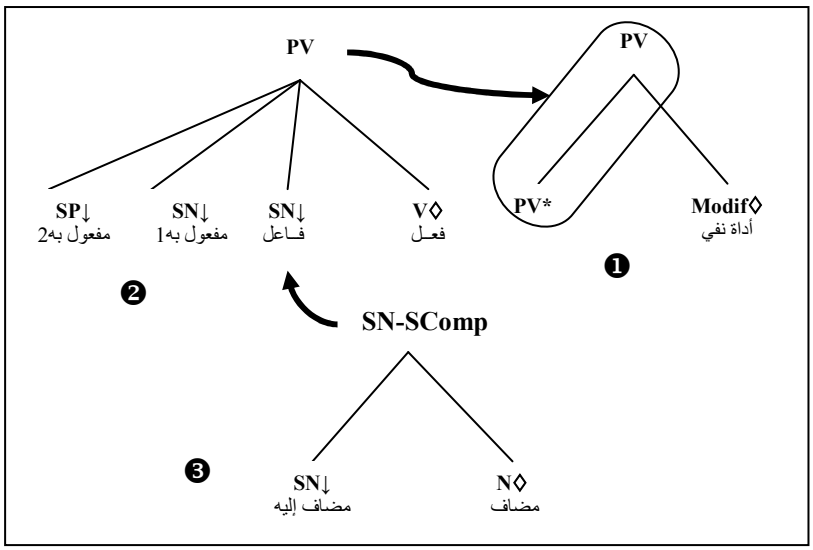


Figure 1 : Premières étapes d'analyse de la phrase P1

Exemple : (P1) لاَ يَقِلُّ دَوْرُ قِطاعِ العَقَارَاتِ أهمّية عَنْ القِطاعَاتِ الاِقْتِصَادِيَّةِ الأُخْرَى (Le secteur foncier n'est pas moins important que les autres secteurs économiques).

Lors de l'analyse de cet exemple, une bonne classification du premier mot لا permettra de récupérer l'arbre élémentaire ❶ représenté sur la figure 1. Cet arbre permet de restreindre les choix de l'arbre élémentaire suivant à ceux représentant seulement des phrases verbales et l'opération de liaison à une adjonction. Ainsi, une classification correcte aboutira en se basant sur les informations morpho-syntaxiques et contextuelles du second mot يَقِلُّ à la structure ❷ de la même figure. Et ainsi de suite pour les autres mots. A chaque fois, le fragment de l'arbre construit à une étape antérieure fournit une information supplémentaire à l'étape courante de la même procédure d'analyse.

# 4 Evaluation de l'apport des attributs discriminants

Afin d'évaluer l'apport des ces attributs dans notre procédure de classification, nous avons utilisé certains modèles standards de classification et ce en faisant appel à la bibliothèque Weka (Witten, Frank, 2005) qui contient les implémentations d'une variété d'algorithmes d'apprentissage automatique standards. Nous avons alors utilisé : les modèles bayésiens NB et LBR pour un apprentissage statistique, le modèle J4.8 (une version améliorée de C4.5) pour

un apprentissage à base d'arbre de décision, l'algorithme IB1 pour un apprentissage à base de cas, les modèles JRip et table de décision comme apprenants à base de règles.

Le corpus d'apprentissage que nous avons utilisé est arboré et est formé d'environ 1050 mots étiquetés par leurs arbres élémentaires correspondants et formant 236 phrases types étiquetées elles mêmes par les arbres de dérivation respectifs. Ces phrases représentent la plupart des structures syntaxiques standards de la langue arabe. L'ensemble de test constitue une partie de celui d'apprentissage et ce dans le cadre d'une validation croisée.

## 4.1 Expériences

Nous avons réalisé des tests en procédant par des ajouts consécutifs des différents attributs d'apprentissage. Nous avons, alors, commencé par un premier test (❶) ne faisant intervenir qu'un ensemble d'attributs morpho-syntaxiques du mot en cours d'analyse. Ensuite nous avons incrémenté cet ensemble par des attributs contextuels. Un second test (❷) a été, ainsi, réalisé avec une fenêtre (-1,+1) et un troisième (❸) a été élaboré avec une fenêtre (-2,+2). Le quatrième test (❹) inclut en plus des attributs morpho-syntaxiques, une fenêtre contextuelle de taille (-1,+1) et des attributs compositionnels. Les critères d'évaluation que nous avons utilisés sont répartis en deux groupes ; à savoir :

- les taux des précisions d'affectation des arbres élémentaires (AEs) aux mots ainsi que celles des opérations choisies pour leur liaison dans le cadre d'une classification élémentaire.

- les taux des précisions d'analyse des phrases dans le cadre d'une classification intégrale.

## 4.2 Exploitation des résultats

Les résultats obtenus sont illustrés dans le tableau 1.

| Classifieurs | | Précisions élémentaires | | Précision d'analyse |
|---|---|---|---|---|
| | | AEs | Opérations | |
| Bayésien | ❶ | 52,33% | 61,95% | 05,00% |
| | ❷ | 57,85% | 68,42% | 06,66% |
| | ❸ | 55,27% | 66,44% | 03,33% |
| | ❹ | 63,75% | 83,21% | 13,33% |
| IB1 | ❶ | 71,34% | 78,31% | 26,66% |
| | ❷ | 98,46% | 96,02% | 81,66% |
| | ❸ | 99,60% | 99,60% | 98,33% |
| | ❹ | 98,45% | 99,04% | 91,66% |
| J48 | ❶ | 59,91% | 70,89% | 06,66% |
| | ❷ | 72,63% | 77,01% | 16,66% |
| | ❸ | 75,29% | 76,87% | 23,33% |
| | ❹ | 76,88% | 95,55% | 30,00% |
| LBR | ❶ | 52,49% | 61,29% | 06,66% |
| | ❷ | 66,01% | 73,35% | 16,66% |
| | ❸ | 68,38% | 75,42% | 20,00% |
| | ❹ | 68,40% | 84,47% | 16,66% |
| JRip | ❶ | 46,50% | 56,39% | 03,33% |
| | ❷ | 53,01% | 63,77% | 11,66% |
| | ❸ | 53,86% | 64,56% | 11,66% |
| | ❹ | 63,09% | 71,14% | 16,66% |
| Table de décision | ❶ | 72,25% | 76,22% | 16,66% |
| | ❷ | 78,49% | 77,79% | 25,00% |
| | ❸ | 78,49% | 77,79% | 25,00% |
| | ❹ | 86,64% | 97,48% | 50,00% |

Tableau 1 : Précisions des classifications : élémentaire et intégrale

Nous avons remarqué en se basant sur ces résultats que la précision de la classification élémentaire ainsi que celle de l'analyse syntaxique intégrale des phrases augmente généralement proportionnellement aux ajouts successifs d'attributs discriminants. L'amélioration est plus élevée pour la classification des opérations de liaison que pour l'affectation des arbres élémentaires. Ceci est dû au fait que le nombre d'opérations à classifier est beaucoup moins élevé que celui des arbres élémentaires à classifier. Aussi, la précision d'analyse des phrases entières a-t-elle augmentée d'un ensemble d'attributs à un autre et ce sauf pour le modèle IB1 qui se base sur l'algorithme des plus proches voisins. De plus, le modèle Bayésien n'a pas amélioré sa performance lors de l'ajout d'une fenêtre contextuelle plus large ; c'est-à-dire (-2,+2) étant donné que la longueur moyenne des phrases du corpus et celles de test est de 4 mots (taille minimale 1 et taille maximale 14). En effet, on pourrait stipuler qu'un contexte plus large, engendrerait des attributs à valeurs nulles ce qui rendrait éparse l'information à extraire.

Par ailleurs, nous avons remarqué que notre choix d'ajouter des informations compositionnelles aux informations standards utilisées dans une analyse syntaxique à base de classification est bénéfique. En effet, nous avons gagné en précision dans la plupart des tests effectués. La discrimination devient alors de plus en plus bonne vu que ces informations restreignent les choix aux nœuds pouvant accepter des dérivations ultérieures. Toutefois, la précision globale de l'analyse est assez faible pour la plupart des modèles d'apprentissage

utilisés pour les tests. Ce qui justifie notre vision de procéder à une classification ensembliste faisant intervenir plusieurs algorithmes d'apprentissage.

# 5 Conclusion et perspectives

Cette étude comparative nous a permis d'avoir une idée sur les impacts des différents attributs d'apprentissage mis en jeu dans la procédure d'analyse syntaxique. Nous avons pu noter que le contexte joue un rôle important dans la discrimination entre structures et nous avons pu justifier expérimentalement l'apport des attributs compositionnels dans ladite procédure d'analyse. Nous avons utilisé plusieurs modèles d'apprentissage afin de rendre plus générales les conclusions tirées. De plus, cette étude comparative nous permettra aussi d'évaluer les performances respectives des différents modèles mis en jeu dans le but d'élaborer un modèle combinatoire d'apprentissage basé sur un apprentissage de qualité.

# Références

BEN FRAJ F., BEN OTHMANE ZRIBI C., BEN AHMED M. (2009). A tool for syntactic tagging an Arabic treebank, *2nd International Conference on Arabic Language Resources and Tools*, Egypt, Cairo.

BEN FRAJ F., BEN OTHMANE ZRIBI C., BEN AHMED M. (2008). Ensemble Classification for Parsing Arabic Texts: A Theoretical Approach, *In Proceedings Artificial Intelligence and Pattern Recognition,* Florida, 51-57.

BEN FRAJ F., BEN OTHMANE ZRIBI C., BEN AHMED M. (2008). ArabTAG: a Tree Adjoining Grammar for Arabic Syntactic Structures, *ACIT2008*, Tunisia, Hammamet.

BEN OTHMANE ZRIBI C. (1998). *De la synthèse lexicographique à la détection et à la correction des graphies fautives arabes*, Thèse de doctorat, Université de Paris XI, Orsay.

COSTA F., LOMBARDO V., FRASCONI P., SODA G. (2001). Wide coverage incremental parsing by learning attachment preferences, *In LNCS 2175, Springer*, 297-307.

DENIS F. & GILLERON R. (2000). *Apprentissage à partir d'exemples*, notes de cours, Université Charles De Gaule, Lille 3.

JOSHI A. (1987) *Introduction to Tree Adjoining Grammar*, in A. Manaster Ramer (ed), The Mathematics of Language, J. Benjamins.s

SAGAE K., LAVIE A. (2005). A classifier-based parser with linear run-time complexity, *In Proceedings of the Ninth International Workshop on Parsing Technologies*, 2005, Vancouver, Canada.

WITTEN I. H., FRANK E. (2005). Data Mining: Practical Machine Learning Tools and Techniques, *2nd Edition*. Morgan Kaufmann.