THÈSE PRESENTÉE A L'UNIVERSITÉ D'ORLÉANS
POUR OBTENIR LE GRADE DE
DOCTEUR DE L'UNIVERSITÉ D'ORLÉANS

Laboratoire d'Informatique Fondamentale d'Orléans
RIADI - Génies Documentiel et Logiciel La Manouba

PAR
Cherifa BEN KHELIL

ÉCOLE DOCTORALE MATHEMATIQUES, INFORMATIQUE, PHYSIQUE THEORIQUE ET INGENIERIE DES SYSTEMES
Discipline : *Informatique*

Construction semi-automatique d'une grammaire d'arbres adjoints pour l'analyse syntaxico-sémantique de l'arabe

Soutenue Publiquement
Le 14 juin 2019 à 9h
Lieu: *Amphithéâtre Herbrand, LIFO, Batiment IIIA - Rue Léonard de Vinci, Orléans.*

MEMBRES DU JURY :

- **Chiraz BEN OTHMANE ZRIBI** Maître de Conférences, ENSI La Manouba
- **Denys DUCHIER** Professeur, Université d'Orléans
- **Claire GARDENT** Directrice de Recherche, CNRS-LORIA UMR 7503
- **Kais HADDAR** Professeur, Université de Sfax
- **Laura KALLMEYER** Professeur, Université de Düsseldorf
- **Yannick PARMENTIER** Maître de Conférences, Université de Lorraine

RÉSUMÉ

Cette thèse traite de la description formelle et du développement d'une grammaire électronique de la langue arabe. Ce travail est un prérequis à la création d'outils de traitement automatique de l'arabe. Cette langue présente de nombreux challenges pour un traitement automatique, en effet l'ordre de mots en arabe est relativement libre, la morphologie y est riche et les diacritiques sont omis dans les textes écrits. Bien que plusieurs travaux de recherche aient abordé certaines de ces problématiques, les ressources électroniques utiles pour le traitement de l'arabe demeurent relativement rares ou encore peu disponibles.

Dans ce travail de thèse, nous nous sommes intéressés à la représentation de la syntaxe (ordre des mots) et du sens de l'arabe standard moderne. Comme système formel de représentation de la langue, nous avons choisi le formalisme des grammaires d'arbres adjoints (Tree Adjoining Grammar). Nous avons ainsi proposé une grammaire d'arbres adjoints électronique de l'arabe nommée « ArabTAGv2 ». Cette ressource réutilise en partie la modélisation pré-existante dans la grammaire définie manuellement «ArabTAG » et l'intègre à une représentation abstraite appelée métagrammaire.

L'expert linguiste peut ainsi décrire la syntaxe et sémantique de la langue avec des outils d'abstraction facilitant la maintenance et l'extension de la grammaire. La grammaire ainsi décrite compte 1074 règles syntaxiques (non lexicalisées) et 27 cadres sémantiques (relations prédicatives). Cette ressource a été évaluée en analysant un corpus issu d'extraits d'un manuel scolaire d'apprentissage de l'arabe.