



**HAL**  
open science

## Analyse des discussions Homme-Homme en arabe

Samira Bendbabis

► **To cite this version:**

Samira Bendbabis. Analyse des discussions Homme-Homme en arabe. Informatique et langage [cs.CL].  
Université de Sfax (Tunisie), 2019. Français. NNT: . tel-02546884

**HAL Id: tel-02546884**

**<https://theses.hal.science/tel-02546884>**

Submitted on 19 Apr 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



# THÈSE DE DOCTORAT

Pour L'obtention du Titre de Docteur en

## Informatique

### *Analyse des discussions Homme-Homme en arabe*

Présentée et soutenue publiquement le 28/12/2019 par :

**Mme. Samira BEN DBABIS BEN ROUINA**

#### Membres du jury :

M. Ahmed HADJ KACEM	Professeur d'Ens. Sup - FSEG - Sfax	Président
Mme. Lamia HADRICH BELGUITH	Professeur d'Ens. Sup - FSEG - Sfax	Directeur de thèse
Mme. Leila BEN AYED	Professeur d'Ens. Sup - ENSI -Tunis	Rapporteur
Mme. Nadia ESSOUSSI	Professeur d'Ens. Sup - ISG -Tunis	Rapporteur
M. Bilel GARGOURI	Maître de Conférences - FSEG - Sfax	Membre
M. Hatem GHORBEL	Professeur d'Ens. Sup -HES-SO-HE-ARC Suisse	Co-Encadrant

Année Universitaire : 2018-2019

## **DEDICACES**

---

*Je dédie cette thèse à :*

*L'âme de mon père qui me manque trop*

*Ma mère et mes frères*

*Mon mari et mes enfants*

*Ma belle-mère et mes belles-sœurs*

## REMERCIEMENTS

---

J'adresse mes sincères remerciements à mon directeur de thèse Mme. Lamia HADRICH BELGUITH, Professeur de l'enseignement supérieur à la Faculté des Sciences Économiques et de Gestion de Sfax (FSEGS) et chef du groupe de recherche ANLP-RG (Arabic Natural Language Processing Research Group) du laboratoire MIRACL, pour la qualité de son encadrement et la pertinence de ses directives. Ses qualités humaines et professionnelles m'ont toujours poussé à surmonter les difficultés rencontrées afin de finaliser ce travail. Je lui suis également reconnaissante pour le temps conséquent qu'elle m'a accordé, sa franchise et sa sympathie. J'ai beaucoup appris à ses côtés et je lui adresse ma profonde gratitude pour son aide considérable tout au long de la préparation de cette thèse.

J'adresse aussi mes vifs remerciements à mon Co-encadrant de thèse Mr. Hatem GHORBEL, Professeur de l'enseignement supérieur à HES-SO-HE-ARC Suisse, pour ses conseils avisés et son écoute qui ont été prépondérants pour la bonne réussite de cette thèse.

Je remercie vivement les membres de jury, pour avoir accepté d'évaluer ce travail. Un grand merci à M. Ahmed HADJ KACEM, Professeur de l'enseignement supérieur à la FSEGS, pour l'honneur qu'il m'a accordée en présidant ce jury. Je remercie également mes deux rapporteurs Mme. Nadia ESSOUSSI, Professeur de l'enseignement supérieur à l'ISG de Tunis et Mme. Leila BEN AYED, Professeur de l'enseignement supérieur à l'ENSIT pour l'intérêt qu'elles ont porté à évaluer mon travail et ce malgré toutes les responsabilités qu'elles assument. Par la même occasion, je remercie M. Bilel GARGOURI, Maître de conférences à la FSEGS, qui m'a fait l'honneur d'examiner ce manuscrit.

Mes remerciements s'adressent aussi à M. Mohamed KALLEL, Maître assistant à la Faculté des Lettres et Sciences Humaines de Sfax, pour son expertise et ses compétences en langue Arabe qui ont favorisé la validation de ce travail.

Je tiens à exprimer aussi tous mes sentiments de respect à tous mes collègues de la Faculté des Sciences Économiques et de Gestion de Sfax qui ont cru en moi et m'ont encouragé tout au long de ce travail.

J'adresse mes plus sincères remerciements à ma famille qui m'a toujours soutenu au cours de la réalisation de cette thèse.

Ces remerciements ne peuvent s'achever, sans une pensée particulière à ma mère. Sa présence et ses encouragements sont pour moi les piliers fondateurs de ce que je suis et de ce que je fais.

# TABLE DES MATIERES

---

TABLE DES FIGURES .....	vii
LISTE DES TABLEAUX .....	viii
<b>INTRODUCTION GENERALE .....</b>	<b>10</b>
<b>CHAPITRE 1: Aperçu général sur la reconnaissance des actes de dialogue .....</b>	<b>13</b>
1. Introduction .....	14
2. De l'acte de langage vers l'acte de dialogue .....	14
3. Reconnaissance des actes de dialogue .....	16
3.1. Segmentation automatique des actes de dialogue .....	17
3.1.1. Apprentissage supervisé .....	17
3.1.1.1. Chaines de Markov Cachées .....	18
3.1.1.2. Réseaux bayésiens .....	20
3.1.1.3. Champs aléatoires conditionnels .....	21
3.1.1.4. Machines à vaste marge .....	22
3.1.1.5. Réseaux de neurones .....	23
3.1.2. Apprentissage semi-supervisé .....	23
3.1.2.1. Auto-Apprentissage .....	24
3.1.2.2. Apprentissage coopératif .....	24
3.1.2.3. Apprentissage Auto-combiné .....	25
3.2. Approche symbolique .....	27
3.1. Tableau récapitulatif .....	29
4. Annotation automatique des actes de dialogue .....	30
4.1. Approche symbolique .....	30
4.2. Approche numérique .....	31
4.2.1. Analyse sémantique latente .....	31
4.2.2. Modèle N-gram .....	31
4.2.3. Chaînes de Markov Cachées .....	32
4.2.4. Réseaux bayésiens .....	33
4.2.5. Réseaux de neurones .....	34
4.2.6. Arbres de décision .....	36
4.2.7. Apprentissage basé transformation .....	37
4.2.8. Apprentissage basé mémoire .....	37
4.2.9. Séparateurs à vaste marge .....	38
4.2.10. Champs aléatoires conditionnels .....	39
5. Conclusion .....	40

<b>CHAPITRE 2: Schémas d'annotation des actes de dialogue</b> .....	<b>41</b>
1. Introduction .....	42
2. Aperçu sur les schémas d'annotation et les corpus annotés.....	42
2.1. Schémas d'annotation.....	42
2.1.1. Le projet MapTask.....	42
2.1.2. Le projet VERBMOBIL.....	43
2.1.3. Le projet DAMSL .....	43
2.1.4. Le projet AMI .....	45
2.1.5. La taxonomie DIT++ .....	45
2.1.6. Le Schéma de (Shala et al., 2010).....	46
2.1.7. Le Schéma de (Elmadany et al., 2014) .....	46
2.2. Corpus annotés .....	47
2.2.1. Le corpus SWITCHBOARD .....	47
2.2.2. Le corpus ICSI-MRDA.....	47
2.2.3. Le projet AMITIES .....	48
2.2.4. Le corpus CallHome .....	48
2.2.5. Le corpus JANA .....	49
2.3. Tableau récapitulatif.....	49
3. Schéma d'annotation SADA .....	49
3.1. Taxonomie générale de la langue Arabe .....	50
3.2. Schéma SADA40.....	51
3.2.1. La gestion des obligations sociales .....	51
3.2.2. La demande .....	52
3.2.3. L'argumentation.....	54
3.2.4. La gestion de dialogue .....	56
3.2.5. La réponse .....	57
3.2.6. La déclaration.....	57
3.2.7. Autres.....	57
3.3. Schéma SADA19.....	60
4. Corpus CARD .....	61
4.1. Corpus CARD 1.0.....	62
4.2. Corpus CARD 1.1.....	62
4.3. Corpus CARD 1.2.....	63
5. Discussion .....	63
6. Conclusion.....	64
<b>CHAPITRE 3: Reconnaissance automatique des actes de dialogue</b> .....	<b>65</b>
1. Introduction .....	66
2. Architecture générale du processus de reconnaissance des actes de dialogue.....	66
3. Apprentissage automatique .....	67
4. Prétraitement .....	70

4.1. Traitement de l'entête.....	71
4.2. Traduction des mots ambigus .....	71
5. Segmentation.....	72
5.1. Segmentation en tours .....	72
5.2. Segmentation en énoncés.....	73
5.2.1. Phase d'apprentissage.....	74
5.2.1.1. Préparation du corpus.....	75
5.2.1.2. Sélection des descripteurs .....	75
5.2.1.3. Apprentissage .....	78
5.2.2. Phase de test.....	78
6. Annotation des actes de dialogue .....	79
6.1. Phase d'apprentissage.....	80
6.1.1. Préparation du corpus .....	81
6.1.2. Sélection des descripteurs .....	81
6.1.2.1. Descripteurs lexicaux .....	81
6.1.2.2. Descripteurs morpho-syntaxiques .....	82
6.1.2.3. Descripteurs structurels .....	83
6.1.2.4. Descripteurs contextuels .....	85
6.1.3. Apprentissage.....	86
6.2. Phase de test.....	88
7. Conclusion.....	89
<b>CHAPITRE 4: Implémentation et évaluation du système RECAARD .....</b>	<b>90</b>
1. Introduction .....	91
2. Présentation générale du système RECAARD.....	91
3. Implémentation du système RECAARD.....	92
3.1. Module de prétraitement.....	92
3.1.1. Traitement de l'entête.....	92
3.1.2. Traduction des mots ambigus .....	93
3.2. Module de segmentation.....	94
3.3. Module d'annotation.....	97
3.3.1. Annotation semi-automatique.....	97
3.3.2. Annotation automatique.....	101
4. Évaluations .....	103
4.1. Mesures d'évaluation.....	103
4.2. Évaluation du module de segmentation.....	104
4.2.1. Expérimentations .....	104
4.2.2. Analyse des cas d'erreurs.....	106
4.3. Évaluation du module d'annotation .....	107
4.3.1. Expérimentations .....	107
4.3.1.1. Résultats de CARD 1.0 .....	108
4.3.1.2. Résultats de CARD 1.1 .....	112

4.3.1.3. Résultats de CARD 1.2 .....	115
4.3.2. Analyse des cas d'erreurs.....	119
5. Conclusion.....	120
<b>CONCLUSION ET PERSPECTIVES .....</b>	<b>121</b>
<b>BIBLIOGRAPHIE .....</b>	<b>124</b>
<b>LISTE DE PUBLICATIONS .....</b>	<b>129</b>

## TABLE DES FIGURES

---

<b>Figure 1.1:</b> Résultats de l'évaluation de WindowDiff.....	28
<b>Figure 2.1:</b> Taxonomie générale de la langue arabe .....	51
<b>Figure 3.1:</b> Démarche suivie pour la reconnaissance des actes de dialogue.....	67
<b>Figure 3.2:</b> modèle graphique de HMM .....	68
<b>Figure 3.3:</b> Modèle graphique de MEMM .....	68
<b>Figure 3.4:</b> Modèle graphique de CRF .....	68
<b>Figure 3.5:</b> Exemple d'entête d'une discussion transcrite .....	71
<b>Figure 3.6:</b> Étapes de la méthode de segmentation proposée.....	72
<b>Figure 3.7:</b> Extrait de la segmentation en tours d'une discussion .....	73
<b>Figure 3.8:</b> Étapes de la phase d'apprentissage .....	74
<b>Figure 3.9:</b> Extrait des mots clés .....	76
<b>Figure 3.10:</b> Mots de question .....	77
<b>Figure 3.11:</b> Liste des signes de ponctuation .....	77
<b>Figure 3.12:</b> Extrait de la liste des POS .....	78
<b>Figure 3.13:</b> Phase de test .....	79
<b>Figure 3.14:</b> Extrait d'une discussion segmentée en énoncés .....	79
<b>Figure 3.15:</b> Phase d'apprentissage.....	80
<b>Figure 3.16:</b> Phase de test .....	88
<b>Figure 3.17:</b> Extrait d'une discussion annotée.....	88
<b>Figure 4.1:</b> Modules du système RECAARD .....	92
<b>Figure 4.2:</b> Exemple d'entête d'une discussion.....	93
<b>Figure 4.3:</b> Extrait des mots ambigus traduits en MSA .....	94
<b>Figure 4.4:</b> Extrait d'une discussion segmentée en tours de parole .....	94
<b>Figure 4.5:</b> Extrait de l'analyse morphologique effectuée par Madamira .....	95
<b>Figure 4.6:</b> Extrait d'un fichier segmenté en mots et en énoncés .....	96
<b>Figure 4.7:</b> Extrait d'une discussion segmentée en énoncés.....	96
<b>Figure 4.8:</b> Interface d'accueil d'ActAAr .....	97
<b>Figure 4.9 :</b> Choix de la discussion à annoter .....	98
<b>Figure 4.10:</b> Choix du schéma d'annotation .....	98
<b>Figure 4.11:</b> Annotation d'un acte de dialogue.....	99
<b>Figure 4.12:</b> Annotation semi-automatique via l'outil ActAAr .....	99
<b>Figure 4.13:</b> Modification d'un acte de dialogue .....	100
<b>Figure 4.14:</b> Recherche d'un acte de dialogue.....	100
<b>Figure 4.15:</b> Statistiques de l'outil ActAAr .....	101
<b>Figure 4.16:</b> Calcul de Kappa via l'outil ActAAr.....	101
<b>Figure 4.17:</b> Sortie finale du système RECAARD .....	102

## LISTE DES TABLEAUX

---

<b>Tableau 1.1:</b> Performance de segmentation utilisant l'information des tours.....	18
<b>Tableau 1.2:</b> Performance de segmentation utilisant l'étiquetage de partie de discours.....	19
<b>Tableau 1.3:</b> Précision des classificateurs pour l'ensemble des tâches de classification .....	20
<b>Tableau 1.4:</b> Comparaison des performances de HMM, MEMM, CRF, MEMM+ et CRF+ .....	21
<b>Tableau 1.5:</b> Comparaison des performances des modèles ME, CRF et Bi-LSTM-CRF .....	23
<b>Tableau 2.1:</b> Les actes de dialogue du projet MAPTASK .....	43
<b>Tableau 2.2:</b> Les 18 actes de dialogue de plus haut niveau de VERBMOBIL .....	43
<b>Tableau 2.3:</b> Les actes de dialogue du projet DAMSL .....	44
<b>Tableau 2.4:</b> Les actes de dialogue du projet AMI .....	45
<b>Tableau 2.5:</b> Les actes de dialogue de (Shala et al., 2010) .....	46
<b>Tableau 2.6:</b> Les actes de dialogue de (Elmadany et al., 2014) .....	47
<b>Tableau 2.7:</b> Statistiques du corpus JANA.....	49
<b>Tableau 2.8:</b> Tableau récapitulatif des corpus annotés.....	49
<b>Tableau 2.9:</b> La gestion des obligations sociales .....	52
<b>Tableau 2.10:</b> Les types de demandes .....	53
<b>Tableau 2.11:</b> La demande .....	53
<b>Tableau 2.12:</b> L'argumentation .....	55
<b>Tableau 2.13:</b> Les types d'arguments .....	55
<b>Tableau 2.14:</b> La gestion de dialogue .....	56
<b>Tableau 2.15:</b> La réponse .....	57
<b>Tableau 2.16:</b> La déclaration.....	57
<b>Tableau 2.17:</b> Autres actes .....	57
<b>Tableau 2.18:</b> Degré d'accord et valeur de Kappa.....	58
<b>Tableau 2.19:</b> Résultats de Kappa par classe .....	59
<b>Tableau 2.20:</b> Schéma d'annotation SADA19 .....	61
<b>Tableau 2.21:</b> Statistiques du corpus CARD 1.0 .....	62
<b>Tableau 2.22:</b> Statistiques du corpus CARD 1.1 .....	62
<b>Tableau 2.23:</b> Statistiques du corpus CARD 1.2 .....	63
<b>Tableau 3.1:</b> Exemples d'expressions ambiguës traduites en MSA .....	72
<b>Tableau 3.2:</b> Exemples de mots clés présents à l'intérieur des actes de dialogue.....	82
<b>Tableau 3.3:</b> Exemples de POS des mots clés.....	83
<b>Tableau 3.4:</b> Exemples d'actes de gestion des obligations sociales.....	84
<b>Tableau 3.5:</b> Exemples d'actes de demande .....	84
<b>Tableau 3.6:</b> Exemples d'actes de gestion de tours de parole .....	85
<b>Tableau 3.7:</b> Exemples de paires d'adjacence .....	86
<b>Tableau 3.8:</b> Exemples de paires d'actes avec même locuteur ou locuteurs différents .....	86
<b>Tableau 3.9:</b> Liste des classes d'actes de dialogue du schéma SADA19 .....	87

<b>Tableau 4.1:</b> Évaluation de la segmentation selon différents modèles d'apprentissage .....	105
<b>Tableau 4.2:</b> Répartition du corpus de discussions CARD.....	105
<b>Tableau 4.3:</b> Résultats d'évaluation de la tâche de segmentation .....	105
<b>Tableau 4.4:</b> Exemple d'une segmentation erronée due à la ponctuation .....	106
<b>Tableau 4.5:</b> Exemple d'une fausse segmentation due aux mots clés .....	107
<b>Tableau 4.6:</b> Résultats d'annotation du corpus CARD 1.0 .....	109
<b>Tableau 4.7:</b> Résultats d'évaluation par acte du corpus CARD 1.0 .....	109
<b>Tableau 4.8:</b> Résultats de la classification du corpus CARD 1.0.....	111
<b>Tableau 4.9 :</b> Résultats de la classification du corpus CARD 1.0 par catégorie .....	111
<b>Tableau 4.10:</b> Résultats d'évaluation de la tâche d'annotation.....	112
<b>Tableau 4.11:</b> Résultats d'annotation du classifieur SMO .....	113
<b>Tableau 4.12:</b> Résultats d'annotation du classifieur NB .....	113
<b>Tableau 4.13:</b> Résultats d'annotation du classifieur J48 .....	113
<b>Tableau 4.14:</b> Résultats d'annotation du classifieur CRF.....	114
<b>Tableau 4.15:</b> Résultats d'évaluation du classifieur CRF avec CARD 1.1 .....	114
<b>Tableau 4.16:</b> Résultats d'évaluation du corpus CARD 1.2 .....	115
<b>Tableau 4.17:</b> Résultats d'évaluation du classifieur CRF avec CARD 1.2 .....	116
<b>Tableau 4.18:</b> Exemples de l'acte de dialogue "Thèse".....	117
<b>Tableau 4.19:</b> Exemple d'une fausse annotation de l'acte de dialogue "Argument" .....	117
<b>Tableau 4.20:</b> Exemples d'arguments de la réalité .....	118
<b>Tableau 4.21:</b> Exemples d'arguments comparaison .....	118
<b>Tableau 4.22:</b> Exemples d'arguments d'autorité.....	118

## INTRODUCTION GENERALE

L'analyse des discussions transcrites reste un défi intéressant dans le domaine de traitement de dialogues Homme-Homme pour différentes langues, et en particulier pour la langue arabe. En effet, les discussions se présentent comme la forme la plus répandue de l'interaction spontanée de la langue parlée. Cette forme de communication s'inspire de l'analyse conversationnelle du discours focalisée sur l'interaction Homme-Homme. Ainsi, les discussions sont conçues comme étant un processus d'échange interactif d'informations qui fait intervenir deux ou plusieurs participants afin de négocier un thème ou une problématique donnée. Les discussions peuvent être virtuelles à travers les forums sur le Web ou réelles dans les réunions de travail, les sessions parlementaires, les émissions télévisées, etc.

Le besoin d'analyser le contenu des discussions s'impose suite à diverses situations. En effet, L'utilisateur peut être une personne qui était absente lors de la discussion et qui veut connaître ce qui s'est passé en termes de participants présents, thématiques discutés, thèses soutenues ou rejetées, points de vue divergents et arguments présentés. D'autre part, l'analyse des discussions peut être faite à haut niveau dans le but de rédiger des comptes rendus ou des résumés de conversations ou encore de suivre et évaluer les attitudes des participants tout au long de la discussion.

Le présent travail s'intègre dans le cadre de l'analyse du discours, plus précisément l'analyse des conversations argumentatives, ce qui consiste à étudier l'aspect communicatif et argumentatif du contenu des discussions en langue arabe parlée spontanée. Cette analyse repose sur la théorie des actes de dialogues qui reflètent le contexte d'énonciation d'une action langagière ce qui permet de décrire la dynamique du dialogue en termes de conflits, d'acceptations, de rejets, d'arguments et de décisions prises.

Pour se faire, nous avons choisi comme corpus de discussions les débats télévisés en langue arabe traitant des sujets politiques d'actualité vu leur richesse en actes argumentatifs explicitant les points de vue des différents participants sur une thématique donnée. Ainsi, chaque intervenant expose un ensemble d'arguments afin de défendre ses idées et convaincre l'audience et les autres invités de l'émission de ses convictions et attitudes.

L'objectif de cette thèse consiste donc à proposer une méthode de reconnaissance des actes de dialogue dans les discussions humaines plus précisément dans les débats télévisés politiques. Le

processus d'identification des actes est fondé sur deux principales étapes à savoir la segmentation des discussions en tours de parole et énoncés et l'annotation des discussions en actes de dialogue.

Pour atteindre les objectifs visés, nous nous sommes orientés vers l'exploitation des méthodes d'apprentissage supervisé pour assurer la reconnaissance automatique des actes de dialogue. Ce choix est justifié par le fait que les algorithmes d'apprentissage ont montré leur efficacité en termes de robustesse et de rapidité dans divers domaines plus précisément dans le domaine de l'analyse de conversations.

Le présent rapport de thèse est organisé autour de quatre chapitres. Le premier chapitre se focalise sur l'état de l'art des travaux réalisés dans le cadre de la reconnaissance automatique des actes de dialogue. Nous détaillons les recherches menées au niveau de chaque étape du processus de reconnaissance à savoir la segmentation et l'annotation des actes de dialogue. Un tableau récapitulatif figure à la fin de chaque tâche illustrant la langue, le corpus utilisé, les caractéristiques ainsi que les résultats d'évaluation de chaque travail de recherche présenté.

Dans le deuxième chapitre, nous mettons l'accent sur trois principaux volets. Dans un premier temps, nous explicitons les différents schémas d'annotation développés souvent dans le cadre de projets d'annotation des actes de dialogue. Nous détaillons les actes de dialogue ainsi que le contexte de réalisation de chaque taxonomie d'actes de dialogue. Étant donné que chaque schéma d'annotation nécessite un corpus d'actes de dialogue pour sa validation, nous présentons dans un deuxième volet les corpus développés dans le cadre de chaque projet d'annotation tout en explicitant le domaine d'application, la langue ainsi que le contexte d'énonciation de chaque corpus annoté. Nous présentons le schéma d'annotation que nous proposons en se basant sur la taxonomie générale de la langue arabe dans un contexte d'énonciation argumentatif à travers les débats politiques télévisés. Nous détaillons ainsi les actes de dialogue proposés dans ce schéma à travers des exemples réels issus de notre corpus de discussions. Pour valider la taxonomie proposée, nous présentons le corpus de débats que nous avons collecté en trois étapes afin d'aboutir à un nombre important d'actes de dialogue (environ 14000 actes) favorisant ainsi l'application de techniques d'apprentissage dans le processus d'identification automatique des actes.

Le troisième chapitre présente la méthode proposée pour la reconnaissance automatique des actes de dialogue. Nous avons exploité les méthodes d'apprentissage supervisé pour assurer le processus de reconnaissance et plus précisément les modèles CRF. Pour se faire nous avons collecté un corpus important de débats télévisés riches en argumentation à partir des transcriptions textuelles des émissions appartenant à des stations télévisées en langue arabe. Nous avons défini aussi un ensemble de caractéristiques linguistiques afin de construire un modèle d'apprentissage capable d'identifier les actes des nouvelles discussions non étiquetés au préalable. Nous détaillons dans ce chapitre le corpus d'apprentissage ainsi que les descripteurs utilisés respectivement dans les tâches de segmentation et d'annotation des actes de dialogues.

Le chapitre quatre s'intéresse à l'implémentation et l'évaluation de la méthode proposée pour la reconnaissance des actes de dialogues. Nous présentons en une première étape l'architecture

générale du système de reconnaissance des actes de dialogue. Ensuite, nous présentons et discutons une série d'évaluations effectuée afin de juger la performance du système proposé. Les évaluations menées sont appliquées sur les deux étapes de segmentation et d'annotation des actes à travers les différentes versions du corpus de débats utilisé. Pour bien valider la performance du système développé, nous procédons à tester d'autres classificateurs afin de confronter les résultats obtenus avec ceux réalisés via l'application des modèles CRF. Une analyse des cas défailants est illustrée à la fin de chaque étape de reconnaissance des actes de dialogue.

Enfin, nous concluons cette thèse en réaffirmant le rôle des modèles CRF dans l'automatisation du processus de reconnaissance des actes de dialogue. Nous dressons le bilan des principaux apports qui caractérisent notre travail et nous proposons quelques perspectives pour les travaux de recherche futurs.

---

## **CHAPITRE 1**

### **Aperçu général sur la reconnaissance des actes de dialogue**

---

## 1. Introduction

L'analyse des discussions humaines s'intègre dans le cadre de l'analyse du discours qui est une approche transdisciplinaire qui traite non seulement les concepts de la linguistique classique tels que la phonologie, la syntaxe, la morphologie et la sémantique mais qui met aussi l'accent sur le contexte d'énonciation et sur les activités du locuteur.

Ainsi, le sujet est considéré comme un acteur sociohistorique agissant par le langage, et la fonction subjective est considérée comme fonction fondamentale de la communication langagière (Maingueneau, 2016).

Le champ de l'analyse du discours est d'autant plus vaste et morcelé qu'on pourrait même parler d'éclatement dans ce domaine. D'où vient la diversité d'approches d'analyse du discours, chacune prenant en considération des aspects particuliers de l'objet discours.

Dans notre contexte de travail, on s'intéresse à l'analyse des débats politiques qui présentent une forme interactive et communicative du discours. L'accent porte principalement à déceler les points de vue des différents participants sur les thématiques posées, les arguments présentés pour défendre leurs idées ainsi que les exemples, les explications et justifications exposés pour convaincre les autres locuteurs ayant des avis contradictoires.

Afin d'atteindre les objectifs cités, l'analyse des débats politiques revient principalement à étudier les intentions des différents locuteurs tout au long de la discussion à travers le concept d'acte de dialogue reflétant l'aspect communicatif et pragmatique du dialogue.

## 2. De l'acte de langage vers l'acte de dialogue

Partant du constat que le langage ne vise pas essentiellement à rapporter un fait, mais permet aussi d'accomplir un acte de nature langagière. Cette conception met en avant les effets que les discours exercent sur les locuteurs-auditeurs. À ce niveau l'attention est focalisée sur la dimension interactive et interactionnelle que toute production langagière présuppose. La pragmatique a donc deux tâches : définir les actes de langage et déterminer quelle proposition est exprimée par un énoncé donné.

Le problème des actes de langage est de trouver les conditions nécessaires et suffisantes pour l'accomplissement normal d'un acte de langage. Ces conditions comportent la présence ou l'absence de certains traits dans le contexte où l'acte de langage est produit (intention du locuteur, savoir, croyance, attentes, effets, etc.).

L'approche pragmatique est tributaire de la théorie des actes de parole développée par (Austin, 1975)et (Searle, 1968). La théorie des actes de langage prend son point de départ dans la conviction que l'unité minimale de la communication humaine, c'est l'accomplissement de certains types d'actes : locutoire, illocutoire et perlocutoire.

- L'acte locutoire :c'est l'acte de dire quelque chose, l'acte d'énonciation proprement dit.
- L'acte illocutoire :c'est l'acte que l'on effectue en disant quelque chose (par exemple : affirmer, conseiller, menacer, promettre, etc.).
- L'acte perlocutoire :c'est l'acte réalisé par le fait de dire quelque chose. Il est caractérisé en termes d'effets réalisés par l'énonciation chez l'interlocuteur (par exemple : effrayer quelqu'un en le menaçant).

En se basant sur la théorie des actes de langage développée par (Austin, 1975), (Searle, 1968)a proposé cinq classes d'actes définies selon le type d'intention illocutoire :

- *Les assertifs* : définis par le fait d'engager la responsabilité du locuteur sur l'existence d'un état de choses. Exemples : affirmer, prédire, avertir, etc.
- *Les commissifs* : obliger le locuteur à faire quelque chose. Exemples : promettre, faire serment, souhaiter, etc.
- *Les directifs* : visent à essayer de faire quelque chose par l'auditeur exprimant la volonté, le souhait, le désir, etc. Exemples : ordonner, supplier, plaider, prier, solliciter, demander, etc.
- *Les déclaratifs* : provoquer un changement par une déclaration.
- *Les expressifs* : permettent d'exprimer un état psychologique. Exemples : s'excuser, féliciter, remercier, etc.

Les actes de dialogue sont définis comme des éléments du comportement communicatif qui permettent de modifier le contexte d'un dialogue. Autrement dit, ce sont les unités fonctionnelles utilisées par le locuteur pour changer le contexte. Ces unités ne correspondent pas aux énoncés en langage naturel ou d'autres instances de la communication d'une manière simple, car en général ce sont des énoncés multifonctionnels.

L'étude des actes de dialogue repose sur la théorie des actes de langage et de la logique illocutoire (Trognon et Batt, 2007) qui associent à chaque énonciation, un contenu propositionnel renvoyant en quelque sorte la sémantique de l'énoncé, et une force illocutoire spécifiant ce que le locuteur désire accomplir par son énonciation. La force illocutoire est elle-

même décomposée en plusieurs caractéristiques dont le but illocutoire constitue la caractéristique la plus importante. Ainsi un acte de dialogue est caractérisé par trois éléments :

- La fonction communicative,
- Le contenu sémantique,
- La forme de l'énoncé.

**Exemple :**

L'acte de dialogue "Est-ce qu'il pleut ?" possède :

- La fonction communicative : "Question Oui/Non",
- Le contenu sémantique : " il pleut ",
- La forme d'énoncé : " Est-ce qu'il pleut ?".

Ces caractéristiques rappellent la définition d'un acte de langage en termes de force illocutoire (Oui/Non) et de contenu propositionnel (il pleut).

Ainsi, interpréter un acte de dialogue signifie attribuer une valeur discursive à cet acte, c'est-à-dire articuler cette action langagière avec l'ensemble des autres actions et déterminer son rôle. Il s'agit de replacer la théorie des actes de langage dans une théorie générale de l'action.

**3. Reconnaissance des actes de dialogue**

La reconnaissance des actes de dialogue consiste à expliciter l'interaction entre les différents participants dans une discussion ainsi que l'échange de messages communicatifs présentés sous forme de demandes, questions, explications, points de vue, etc.

L'identification des actes de dialogue est principalement fondée sur deux étapes primordiales à savoir la segmentation en unités fonctionnelles ainsi que l'annotation en actes de dialogue.

L'étape de segmentation consiste à décomposer la discussion en tours de parole ; chaque tour est segmenté par la suite en énoncés. Dans l'étape d'annotation, à chaque énoncé est assigné un acte de dialogue exprimant l'intention du locuteur dans le contexte d'énonciation.

Plusieurs chercheurs considèrent qu'un tour de parole est formé d'un seul énoncé ((Bahou et al., 2008), (Shala et al., 2010), (Graja et al., 2013), (Lhioui et al., 2013) et (Hijawi et al., 2013)). Ainsi la reconnaissance des actes de dialogue consiste à étiqueter les tours de parole avec les actes correspondants.

D'autres travaux regroupent les deux étapes en une seule c'est-à-dire la discussion est segmentée et annotée simultanément. La combinaison des deux tâches a montré son efficacité en terme d'exactitude. En effet, (Tavafi et al., 2013) ont obtenu une valeur d'exactitude égale à 80,5% en

combinant les techniques d'apprentissage SVM et HMM appliquées sur le corpus ICSI-MRDA (Shriberg et al., 2004).

Dans notre contexte, les expérimentations ont montré des résultats plus pertinents en séparant les deux étapes. En effet, la performance de l'identification des actes de dialogue (en terme de F-mesure) se dégrade à 52% en joignant la segmentation à l'annotation alors que ce taux atteint 63,4% en suivant un processus séquentiel pour ces deux tâches.

Cette baisse peut être expliquée par l'importance de bien délimiter les segments avant de les annoter. Ainsi les erreurs de segmentation influencent directement sur l'attribution des actes de dialogue appropriés. De même la définition de descripteurs spécifiques à chaque tâche améliore fortement les résultats d'apprentissage attendus. Par exemple, les marques de ponctuation ont une pondération importante dans l'étape de segmentation que dans l'étape d'annotation.

Nous détaillons dans ce qui suit l'état de l'art des travaux effectués pour la reconnaissance des actes de dialogue via les deux étapes de segmentation et d'annotation.

### 3.1. Segmentation automatique des actes de dialogue

La segmentation des discussions Homme-Homme est une tâche clé dans le processus d'analyse automatique du discours. Ainsi, cette étape s'avère très importante en raison de l'impossibilité d'atteindre les phases supérieures telles que l'annotation, l'analyse syntaxique et l'analyse sémantique.

Dans la littérature nous distinguons deux principales approches à savoir l'approche par extraction appelée *numérique* et l'approche par compréhension dite *symbolique*.

L'approche numérique est basée principalement sur l'apprentissage automatique alors que l'approche symbolique est fondée sur des méthodes linguistiques à savoir l'analyse lexicale, morphologique, syntaxique et sémantique.

Dans cette section, nous présentons l'état de l'art des travaux de segmentation automatique des discussions en énoncés tout en mettant l'accent sur les caractéristiques de chaque méthode.

#### 3.1.1. Apprentissage supervisé

La plupart des méthodes numériques adoptent la technique d'apprentissage supervisé. Cette technique se base sur deux phases : la phase d'apprentissage et la phase de test. Ces deux phases nécessitent des corpus étiquetés. L'étiquetage d'une large base des dialogues est très coûteux ce qui représente l'inconvénient majeur de cette technique.

Nous présentons dans cette section les classificateurs qui implémentent la méthode d'apprentissage supervisé.

### 3.1.1.1. Chaines de Markov Cachées

Le modèle de Markov caché (Hidden Markov Model : HMM) est une fondation formelle pour la construction des modèles probabilistes pour un problème d'étiquetage de séquences linéaires (Blasiak et al., 2011). Les modèles de Markov cachés sont exploités dans plusieurs domaines. Dans notre contexte de travail, ce modèle est utilisé pour trouver les limites de segments les plus probables.

(Stolcke et al., 2000) est l'une des références les plus connues dans le domaine de segmentation automatique du dialogue à travers l'exploitation des chaînes de Markov cachées. Leur travail consiste à former un *modèle de langage statistique* de variété n-gram pour modéliser la distribution des mots et les limites de segments. Ce modèle de langage est utilisé pour prédire les localisations des limites des segments les plus probables. Pendant le test, le modèle est exécuté comme un *modèle de segmentation caché* en faisant l'hypothèse des limites du segment entre deux mots et en calculant implicitement les probabilités de toutes les segmentations possibles. À chaque mot peut être associé l'état <S> s'il présente la limite du segment ou l'état <NO-S> s'il est à l'intérieur du segment en calculant les deux distributions de probabilité :  $P_S$  et  $P_{NO-S}$ .

Leur corpus d'apprentissage est composé de 1,4 million de mots, à partir des transcriptions annotées par le projet UPenn Treebank (Meteer et al., 1995) comprenant 193000 segments dont 10000 mots et 1300 segments sont utilisés pour le test.

Pour évaluer leur expérience, (Stolcke et al., 2000) ont ajouté d'autres informations à savoir :

- Les informations des tours de parole : Ils ont incorporé ces informations dans le modèle de segmentation en ajoutant des balises spéciales au niveau des limites des tours dans le corpus d'entraînement et de test. La performance de ce modèle est résumée dans le tableau 1.1.

	Rappel	Précision	Taux d'erreur
BaseLine	70,2 %	60,7%	53,1%
Marquage des tours	76,9%	66,9%	44,9%

Tableau 1.1: Performance de segmentation utilisant l'information des tours

Ainsi, l'ajout de l'information des tours améliore la performance du modèle de segmentation dans tous les indicateurs.

- Les informations des parties des discours (POS) : ils ont ajouté l'étiquette de la partie de discours dans le modèle de segmentation utilisant l'information des limites de tours. La performance de ce modèle est illustrée dans le tableau 1.2.

	Rappel	Précision	Taux d'erreur
Mot	76,9 %	66,9%	44,9%
POS	79,6%	73,5%	39,9%

Tableau 1.2: Performance de segmentation utilisant l'étiquetage de partie de discours

L'utilisation des étiquettes de POS a aussi augmenté la performance du modèle de segmentation qui est démontré par l'amélioration de tous les indicateurs comme indiqué dans le tableau 1.2.

Dans ce contexte, (Ivanovic, 2005b) s'est inspiré du travail de (Stolcke et al., 2000) en exploitant les chaînes de Markov dans la segmentation automatique des messages instantanées en énoncés. Il a utilisé le corpus de (Ivanovic, 2005a) à travers des sessions de messagerie instantanée en ligne. Ce corpus est formé par 550 énoncés contenant 6500 mots avec 10 mots/message en moyenne. Il a consacré huit sessions de chat pour l'apprentissage et une session pour le test.

Ainsi (Ivanovic, 2005b) a expérimenté le modèle HMM avec trois versions basées sur les séquences de lemmes, les étiquettes de POS et les mots début des segments.

D'abord, il a formé un modèle de langage statistique pour prédire les endroits les plus probables de limites d'énoncés pour chaque message individuel. Les débuts des segments dans les données d'entraînement sont marqués avec la balise <S>. Ainsi chaque élément est dans l'un des deux états :S ou NO-S selon la présence ou non d'une balise <S>.

L'algorithme de Viterbi (Fomey et David, 2005) correspondant est ensuite utilisé pour trouver la séquence la plus probable des états S et NO-S étant donné les mots lemmatisés. Puisque la tâche de segmentation est essentiellement de nature subjective, les mesures de rappel et précision habituellement utilisés peuvent considérer une alternative de segmentation comme erreur. Donc, les résultats de segmentation obtenus par (Ivanovic, 2005b) ont été évalués en utilisant la métrique de WindowDiff. Cette mesure définit une taille de fenêtre  $k$  qui est typiquement égale à la moitié de la longueur moyenne des segments dans le corpus.

Cette taille de fenêtre  $k$  glisse ensuite sur la segmentation supposée des données et compare le marquage segment et non-segment avec les données de référence. Si le nombre de segments supposés et de segments de référence dans la fenêtre de taille  $k$  se différencie, alors le compteur est incrémenté et la fenêtre continue à la position suivante. Le score final est ensuite divisé par le

nombre des *scans* effectués. Le meilleur résultat obtenu par la valeur la plus basse de WindowDiff. Il a ajusté la valeur de k allant de 1 à 20 pour chaque message. Donc, il a obtenu les résultats suivants avec k =15 :

HMM avec les étiquettes de partie de discours : WindowDiff = 0,25.

HMM avec les mots de début des segments : WindowDiff = 0,21.

HMM avec les lemmes des mots : WindowDiff = 0,17.

### **3.1.1.2. Réseaux bayésiens**

Les réseaux bayésiens sont des modèles graphiques probabilistes initiés par (Pearl et Russel, 2002). Ils se sont révélés des outils très pratiques pour la représentation de connaissances incertaines et le raisonnement à partir d'informations incomplètes. Ainsi, ils ont été utilisés dans plusieurs domaines et plus précisément dans l'identification des limites des segments des dialogues humains.

Le travail de (Geertzen et al., 2007) exploite les réseaux bayésiens pour la segmentation automatique des conversations en énoncés ou en segments fonctionnels. Ils ont utilisé le corpus DIAMOND (Geertzen et al., 2004) qui est formé de dialogues Homme-Homme en néerlandais. Ce corpus est constitué des dialogues transcrits et contient 952 énoncés représentant 1408 segments fonctionnels.

Les descripteurs utilisés par ce travail sont liés à l'historique du dialogue (les étiquettes des 4 tours précédents), la prosodie (le ton minimum, maximum et moyen, l'énergie et la voix) et l'occurrence des mots. Ils ont utilisé la validation croisée pour l'apprentissage et le test.

Le tableau 1.3 donne une vue d'ensemble de la performance des classificateurs testés sur les fonctions de communication par dimension. En effet, les réseaux bayésiens donnent des taux d'exactitude meilleurs que ceux obtenus par le Baseline (Exactitude=48%). Aussi, la dimension *Gestion de temps* donne une meilleure exactitude de 99,2%.

<b>Tâche de classification</b>	<b>BaseLine</b>	<b>Réseaux Bayésiens</b>
Etiquette Dimension	38%	69,5%
Gestion de tâches	66,8%	71,2%
Auto-Feedback	77,9%	86%
Début de tour	93,2%	92,9%
Fermeture de tour	58,9%	85,1%
Gestion de temps	69,7%	99,2%
Etiquette fonctionnelle	25,7%	48%

*Tableau 1.3 : Exactitude des classificateurs pour l'ensemble des tâches de classification*

**3.1.1.3. Champs aléatoires conditionnels**

Les champs aléatoires conditionnels (*Conditional Random Fields :CRF*) sont des modèles discriminatifs conditionnels non orientés. Ils sont souvent utilisés dans différentes tâches d'étiquetages des données séquentielles.

Les modèles CRF sont introduits par (Lafferty et al., 2001) pour remédier aux insuffisances des autres modèles discriminatifs de Markov à savoir les modèles MEMM (Maximum Entropy Markov Models). CRF évite également la restriction fondamentale de l'entropie maximale des modèles de Markov (MEMM) et des autres modèles de Markov discriminatifs fondés sur des modèles graphiques dirigés.

(Lafferty et al., 2001) ont démontré l'excellence de CRF dans la tâche d'étiquetage de la partie de discours (POS) de séquences des mots par rapport à HMM de premier ordre et à MEMM en se basant sur le corpus du projet Penn Treebank (Meteer et al., 1995) avec 50% de ce corpus pour l'entraînement et 50% pour le test.

Les résultats obtenus par (Lafferty et al., 2001) sont illustrés dans le tableau 1.4 contenant les taux d'erreur dûs à l'utilisation de CRF, MEMM et HMM de premier ordre dans la tâche d'étiquetage de parties de discours (POS) sans l'utilisation des descripteurs et puis avec l'utilisation des descripteurs orthographiques à savoir si le mot commence par un numéro ou une lettre majuscule, si elle contient un trait d'union, ou si elle se termine par un des suffixes suivants: *-ing, -ogy, -ed, -s, -ly, -ion, -tion, -ity* et *-ies*.

Ils ont remarqué que les MEMMs et les CRFs performant de manière significative avec l'utilisation de ces descripteurs en constatant que le taux d'erreur global a été réduit d'environ 25% ainsi que le taux d'erreur des mots hors vocabulaire (*oov* : out-of-vocabulary) a été réduit d'environ 50%.

Modèle	Taux d'erreur	Taux d'erreur ( <i>oov</i> )
HMM	5,69%	45,99%
MEMM	6,37%	54,61%
CRF	5,55%	48,05%
MEMM <sup>+</sup>	4,81%	26,99%
CRF <sup>+</sup>	4,27%	23,76%

*Tableau 1.4 : Comparaison des performances de HMM, MEMM, CRF, MEMM+ et CRF+*

Ainsi grâce à la robustesse du CRF, (Silvia et al., 2011) utilisent le modèle discriminatif des champs aléatoires conditionnels dans leur travail qui consiste à la segmentation simultanée des

dialogues humains en tours puis à l'intra-segmentation de ces tours en énoncés et enfin à la classification de chaque segment en acte de dialogue.

(Silvia et al., 2011) utilisent deux corpus : SWITCHBOARD qui est constitué de 1155 dialogues humains en anglais et LUNA formé de 94 dialogues humains en italien. Pour expérimenter ce modèle, ils combinent plusieurs descripteurs tels que la liste des mots, la partie de discours et le modèle de CoNLL 2000. Les résultats obtenus montrent que le modèle de CoNLL 2000 donne la meilleure valeur d'exactitude au niveau mot ( $Accuracy_w = 70,9$ ) et au niveau tour ( $Accuracy_T = 58,8\%$ ) pour la tâche de segmentation.

#### 3.1.1.4. Machines à vaste marge

Les machines à vaste marge (Support Vector Machine : SVM) sont des méthodes de classification par apprentissage supervisé qui exploitent les concepts relatifs à la théorie de l'apprentissage statistique et à la théorie des bornes. SVM repose sur l'existence d'un classificateur linéaire dans un espace approprié. Il fait appel à un jeu de données d'apprentissage pour apprendre les paramètres du modèle. Ainsi SVM est basée sur l'utilisation de fonctions noyau (kernel) qui permettent une séparation optimale des données.

Cette méthode est souvent utilisée pour résoudre des problèmes de classification binaire. Dans cette perspective, plusieurs travaux ont exploité les modèles discriminatifs SVM pour identifier les limites de segments de dialogue. En effet le problème de segmentation revient à classifier chaque mot du segment soit B (Begin) indiquant le début d'un segment soit I (Inside) présentant un mot à l'intérieur du segment.

(Elmadany et al., 2015) ont exploité SVM pour segmenter les énoncés des dialogues issus de messages téléphoniques instantanés et dialogues spontanés portant sur des renseignements aériens et bancaires. Ils ont utilisé le corpus JANA formé de 82 dialogues contenant 4725 énoncés en dialecte égyptien. Ils ont intégré des caractéristiques d'ordre lexical, morphologique et contextuel. Ils ont ainsi obtenu des résultats encourageants (F-mesure=90,74%).

Récemment, (Ngo et al., 2017) ont utilisé les modèles CRF pour découper les dialogues vietnamiens en actes de dialogue. Ils ont exploité deux types de corpus : un premier corpus formé de messages issues de réseaux sociaux inclus dans 280 dialogues et un deuxième corpus de conversations téléphoniques. Ils ont obtenu des résultats d'évaluation prometteurs pour les messages (F-mesure=74,51%) et pour les communications téléphoniques (F-mesure=57,27%).

3.1.1.5. Réseaux de neurones

Les réseaux neuronaux sont des modèles connexionnistes qui font partie des technologies de l'apprentissage profond (*Deep Learning*).

Un réseau neuronal s'inspire du fonctionnement des neurones biologiques et prend corps dans un ordinateur sous forme d'un algorithme. Il est en général composé d'une succession de couches ; chaque couche (i) est composée de  $N_i$  neurones, prenant leurs entrées sur les  $N_{i-1}$  neurones de la couche précédente.

(Ngo et al., 2017) ont introduit les modèles basés sur l'apprentissage profond pour la segmentation des énoncés en vietnamien. Ils ont intégré les réseaux de neurones récurrents bidirectionnels basés mémoire (Bi-LSTM : Bidirectional Long Short-Term Memory) combinés avec les modèles CRF. Ils ont utilisé comme corpus 4587 messages des réseaux sociaux inclus dans 280 dialogues portant sur des thématiques variées. Ils ont aussi exploité 22 conversations téléphoniques contenant 3500 segments fonctionnels. Les résultats d'évaluation montrent la performance de l'apprentissage profond par rapport aux méthodes d'apprentissage classique testées à savoir l'entropie maximale (ME : Maximum entropy) et les modèles CRF. En effet, le modèle Bi-LSTM-CRF a donné une valeur de F-mesure=90,42% pour les messages issus de réseaux sociaux et 73,26% pour le corpus de conversations téléphoniques. La performance des résultats du corpus formé de messages textuels par rapport aux conversations téléphoniques est expliquée par le fait que les tours de parole inclus dans les messages sont plus concis et moins ambigus. Les conversations téléphoniques présentent souvent des disfluences telles les hésitations, les répétitions et la discontinuité dans l'enchaînement des tours de parole.

Le tableau 1.5 illustre l'étude comparative entre les résultats de détection des segments fonctionnels utilisant les modèles ME, CRF et Bi-LSTM-CRF.

Modèle	Messages			Conversations téléphoniques		
	Précision	Rappel	F-mesure	Rappel	Précision	F-mesure
ME	57,38, %	57,33%	57,34%	61,88%	56,23%	58,92%
CRF	83,8%	67,08%	74,51%	66,82%	50,18%	57,27%
Bi-LSTM-CRF	91,3%	89,56%	<b>90,42%</b>	77,42%	60,47%	<b>73,26%</b>

Tableau 1.5 : Comparaison des performances des modèles ME, CRF et Bi-LSTM-CRF

3.1.2. Apprentissage semi-supervisé

La technique d'apprentissage semi-supervisé est la combinaison des deux techniques d'apprentissage :

- Supervisé : qui utilise seulement des données étiquetées.
- Non supervisé : qui utilise des données non étiquetées.

L'apprentissage semi-supervisé suppose qu'il y a une petite quantité de données étiquetées et une grande quantité de données non étiquetées. Elle dispose seulement d'une phase d'apprentissage et d'une phase de test des résultats générés par le classificateur sans aucune phase de validation. Donc l'avantage le plus important de cette technique est qu'elle offre un gain de temps en terme d'étiquetage des données d'entraînement.

(Dalva et al., 2018) ont exploré la technique d'apprentissage semi-supervisé pour la segmentation automatique des dialogues en actes de dialogue. Dans ce contexte, ils appliquent des algorithmes qui implémentent cette technique à savoir l'auto-apprentissage, l'apprentissage coopératif et l'apprentissage auto-combiné.

### ***3.1.2.1. Auto-Apprentissage***

C'est la méthode d'apprentissage semi-supervisé la plus populaire. Afin d'auto-apprendre, le modèle donné estime les classes pour la partie des données non marquées. Ensuite, les exemples classés automatiquement sont ajoutés à l'ensemble d'entraînement, le modèle est recyclé et l'ensemble du processus est itéré (Clark et al., 2018).

Pour éliminer le bruit introduit par les échantillons faussement classés, seuls ceux qui sont classés avec un niveau de confiance élevé peuvent être exploités. De même, ceux qui sont classés avec une très grande confiance peuvent aussi être ignorés pour éviter la tendance de facilité de classement.

L'auto-apprentissage a été appliqué à un certain nombre de tâches de Traitement Automatique du Langage Naturel (TALN), comme le marquage de la partie de discours, la désambiguïsation, et l'analyse syntaxique. En auto-formation l'ensemble des caractéristiques est considéré comme une vue unique. Un seul modèle étiquette automatiquement les exemples par lui-même. Pour réduire le bruit de classification, (Dalva et al., 2018) utilisent un seuil  $\theta$  pour sélectionner les exemples les plus confiants à chaque itération.

### ***3.1.2.2. Apprentissage coopératif***

Contrairement au concept traditionnel de la machine d'apprentissage avec une seule vue, cette méthode implémente l'approche multi-vues c'est-à-dire l'ensemble des descripteurs consiste en deux sous-ensembles ou plus des différents descripteurs.

La formation coopérative est l'une des approches les plus efficaces de l'apprentissage automatique semi-supervisé basée sur l'apprentissage multi-vues en associant une hypothèse

pour chaque vue et en ajoutant à l'ensemble d'entraînement les prédictions les plus confiantes faites sur les exemples non étiquetés de façon répétée.

Cette méthode est très efficace dans plusieurs tâches de classification à savoir la classification des pages web, la désambiguïsation des sens des mots et la reconnaissance des entités nommées. Elle améliore les performances de l'algorithme d'apprentissage supervisé par l'incorporation d'une large quantité des données non étiquetées dans le corpus d'apprentissage. L'algorithme fonctionne avec la génération de deux classificateurs ou plus entraînés avec différentes vues des entrées étiquetées qui sont utilisés séparément pour étiqueter les données non étiquetées. Les exemples étiquetés automatiquement seront ensuite ajoutés aux données étiquetées manuellement. Ce processus peut être répété pour plusieurs itérations.

(Dalva et al., 2018) appliquent cette méthode en utilisant les descripteurs prosodiques et lexicaux comme deux vues séparées des données et il utilise une version étendue de l'algorithme de base de l'apprentissage coopératif. L'approche d'apprentissage coopératif de (Dalva et al., 2018)est constituée de plusieurs étapes. Dans la première étape, Ils entraînent deux modèles séparés  $M_{pro}$  (utilisant seulement des caractéristiques prosodiques) et  $M_{lex}$  (utilisant seulement des caractéristiques lexicales) qui sont basés sur des petites quantités de données étiquetées manuellement ( $L$ ). Ensuite, ils estiment les limites des segments pour les portions des données non encore étiquetées ( $U$ ) en utilisant ces deux modèles. Les exemples  $x_i$  sont triés en fonction de leurs scores de confiance pour les deux cas. À ce stade, ils ont essayé différents exemples de mécanismes de sélection pour monter avec l'ensemble des exemples obtenus par les deux modèles.

### 3.1.2.3. Apprentissage Auto-combiné

Dans cette approche, (Dalva et al., 2018) considèrent les exemples qui ont les scores de confiance les plus élevés obtenus à partir de l'auto-apprentissage des modèles prosodiques et lexicaux individuellement dans chaque itération.

D'abord, comme un processus d'initialisation, les données manuellement étiquetées  $L$  (qui contiennent à la fois les caractéristiques prosodiques et lexicales) sont séparées en deux parties selon leurs classes respectives  $L_{pro}$  et  $L_{lex}$ .

Ensuite, ils ont formé les modèles  $M_{pro}$  et  $M_{lex}$  en utilisant cette petite quantité de données étiquetées manuellement. Ils obtiennent des scores de confiance estimés  $f M_{pro}(x_i, pro)$  et  $f M_{lex}(x_i, lex)$  des grandes données non étiquetées  $U_{pro}$  et  $U_{lex}$  par l'utilisation de ces deux modèles formés.

Les exemples ayant les plus hauts scores de confiance sont déterminés par le tri des scores de confiance dans l'ordre décroissant et en fixant différents seuils.

L'ensemble du processus est réitéré jusqu'à ce que les ensembles non étiquetés soient complètement étiquetés ou les modèles ont atteint le seuil maximal d'amélioration. A chaque itération, les modèles prosodiques et lexicaux sont formés individuellement (auto-formation basée sur une vue).

L'évaluation des performances est réalisée en testant les ensembles de données dans chaque itération après l'obtention de nouvelles données étiquetées qui contiennent à la fois les caractéristiques prosodiques et lexicales. A ce stade, ils forment un modèle à chaque itération en utilisant les données marquées dans le corpus de test. Toutes les expériences sont réalisées en utilisant la transcription manuelle pour éviter le bruit introduit par le système de reconnaissance vocale.

(Dalva et al., 2018) ont réalisé des expériences en utilisant différentes tailles de données étiquetées manuellement initialement et puis ils ont comparé les différentes méthodes semi-supervisées d'apprentissage. Ils ont utilisé le corpus des réunions ICSI-MRDA (Shriberg et al., 2004) contenant 72 heures des dialogues télévisés. Les conversations sont divisées en 51 pour l'apprentissage (539 K mots), 11 pour la validation (110 K mots) et 11 conversations pour le test (102 K mots). Dans leurs expériences, ils ont utilisé des descripteurs prosodiques à savoir la durée de la pause entre deux mots, les mesures de la hauteur et l'énergie de la voix du parlant. De plus, ils ont utilisé des descripteurs lexicaux à savoir les n-grams composés du mot après la limite du segment en question et les deux mots précédents.

(Dalva et al., 2018) ont utilisé le classificateur discriminatif "*Boosting*" implémenté avec l'outil *BoosTexter*. Pour toutes les expériences, ils ont itéré le Boosting 500 fois. L'évaluation des différentes stratégies de l'approche semi-supervisée est effectuée avec les métriques F-mesure et le taux d'erreur NIST. Les résultats ont confirmé les points suivants :

- Avec un petit nombre de données étiquetées manuellement (1000):

La stratégie "*Auto-combiné*" donne une meilleure performance dans la première itération (F-mesure=71,19%, Taux d'erreur=51,29%) alors qu'après un certain nombre d'itérations la stratégie "*Désaccord*" donne la meilleure performance (F-mesure=74,16%, Taux d'erreur=48,34%).

- Avec un moyen nombre de données étiquetées manuellement (3000) :

La stratégie "Auto-combiné" donne une meilleure valeur de F-mesure dans la première itération (F-mesure=75,55%) mais le plus bas taux d'erreur est obtenu par la stratégie "Désaccord" (NIST=45,59%). Après un certain nombre optimal d'itérations la stratégie "Désaccord" donne une meilleure performance (F-mesure=77,22% , Taux d'erreur=43,75%).

- Avec un nombre important de données étiquetées manuellement (6000):

La stratégie "Auto-combiné" donne la meilleure performance dans la première itération (F-mesure=77,19%, Taux d'erreur NIST=43,18%) et après un certain nombre d'itérations une valeur de F-mesure=77,87% et un taux d'erreur NIST=42,28%.

### 3.2. Approche symbolique

Dans les dernières années, plusieurs chercheurs en TALN ont montré l'utilité d'exploiter l'approche symbolique dans la segmentation automatique des dialogues Homme-Homme en énoncés. En effet, cette approche est basée sur l'analyse linguistique du discours et de sa structure à travers la représentation formelle des phrases.

Dans ce contexte, (Ivanovic, 2005b) a utilisé la méthode d'arbre syntaxique dans le même travail déjà présenté dans la section 3.1.1.1 qui a comme objectif la segmentation automatique des messages instantanés en énoncés. En effet, il a fait une étude comparative des performances des chaînes de Markov cachées et de l'arbre d'analyse syntaxique.

(Ivanovic, 2005b) a supposé que les limites d'énoncés se produisent uniquement lors des majeures limites syntaxiques. Ceci est similaire à son principe de l'utilisation de segments comme décrit dans la méthode de HMM où il émet l'hypothèse que la limite de segment existe avant chaque jeton. Cependant, la notion d'un jeton change de la représentation des segments à la représentation des sous-arbres dans un arbre syntaxique.

Dans ce contexte, un jeton représente plusieurs mots et les segments d'énoncés ne se présentent qu'entre les jetons. Cette méthode réduit considérablement la possibilité d'obtenir des limites de segment faux-positifs. Mais, si les arbres de dérivation ne sont pas corrects, cette technique aura l'effet inverse. Les segments d'énoncés des données utilisés par (Ivanovic, 2005b) se produisent toujours à une profondeur maximale de 2 nœuds depuis la racine de l'arbre d'analyse. L'utilisation de cette limite de profondeur, permet de construire d'abord un tableau des "coupures" possibles à travers l'arbre. Ces coupures ou ces analyses appropriées telles que décrites dans (Chomsky, 1965) contiennent toutes les combinaisons des sous-arbres. Ce qui entraîne une séquence C de nœuds  $C = t_1, t_2, t_3, \dots, t_n$ . Où chaque combinaison  $t_i$  est une

séquence des nœuds d'arbre de sorte que  $t_i = t_1, t_2, t_3, \dots, t_n$ . La feuille de chaque nœud d'arbre  $t_i$  représente un énoncé possible.

Ensuite, il calcule les actes de dialogue les plus probables pour les feuilles (mots) de chaque nœud. L'acte de dialogue le plus probable correspondant est stocké dans le nœud  $t_i$ .

Puis, il calcule la probabilité d'une séquence correcte des énoncés en utilisant les formules suivantes :

$$\langle t^*, d^* \rangle = \arg \max \prod P(d_i | t_i, d_{i-1}) \quad (1)$$

$$P(d_i | t_i, d_{i-1}) = P(d_i | t_i, d_{i-1}) \prod P(v | d_i) \quad (2)$$

Où  $t^*$  est la meilleure combinaison de nœuds (segments),  $C$  est l'ensemble des analyses appropriées,  $P(d_i | t_i)$  est la probabilité du nœud  $t_i \in t$  d'être un acte de dialogue  $d$  en fonction de ses feuilles (mots),  $d_{i-1}$  est l'acte de dialogue précédent et  $v$  est un mot dans le nœud  $t_i$ .

L'utilisation de cette méthode a pour effet d'évaluer en même temps les tâches de classification et de segmentation, en prenant la combinaison la plus probable. La méthode de classification renvoie la probabilité la plus élevée de tous les actes de dialogue ce qui donne les mots dans le nœud  $n$  en utilisant la méthode des réseaux bayésiens. Elle renvoie aussi l'acte de dialogue correspondant qui est ensuite stocké dans le nœud  $n$ . Les corpus d'apprentissage et de test des deux méthodes (HMM et Arbre d'analyse syntaxique) sont décrits dans la section 3.1.1.1. Ainsi, leur évaluation est réalisée à travers la métrique WindowDiff.

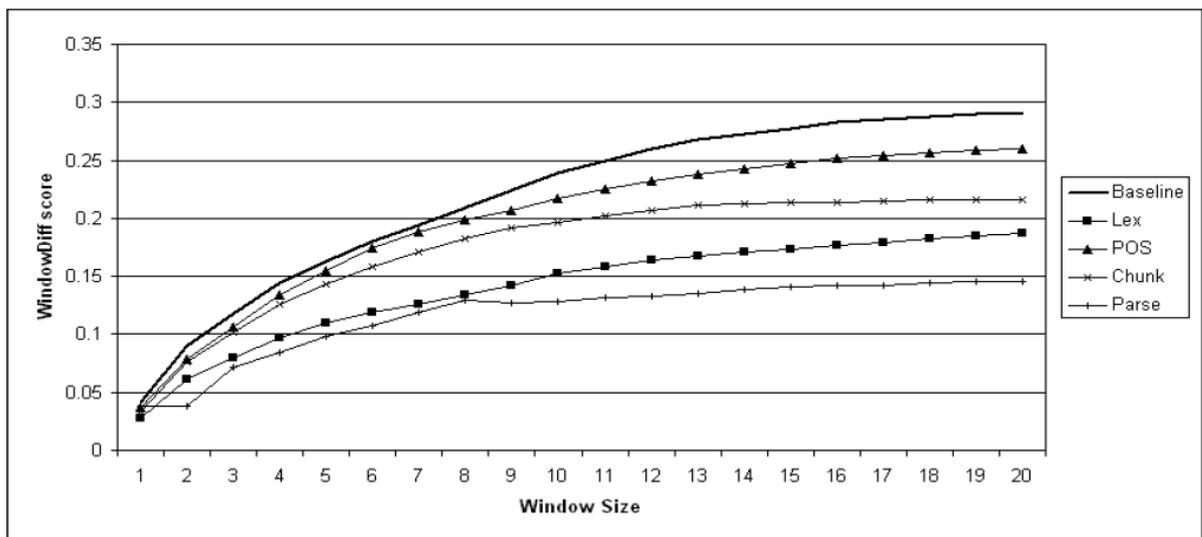


Figure 1.1: Résultats de l'évaluation de WindowDiff

(Ivanovic, 2005b) a constaté après son étude que l'arbre d'analyse syntaxique est meilleur que les chaînes de Markov cachées dans la tâche de segmentation automatique des messages instantanés en énoncés car elle a toujours donné la plus petite valeur de WindowDiff (pour  $k$  de 1 à 20). Les résultats d'évaluation des deux méthodes utilisées sont illustrés dans la figure 1.1.

3.1. Tableau récapitulatif

Référence	Corpus	Langue	Descripteurs	méthode utilisée	Résultats
Geertzen et al., 2007	Corpus DIAMOND de 952 énoncés.	néerlandais	-Descripteurs lexicaux -Descripteurs prosodiques	Réseaux bayésiens	<i>Accuracy</i> = 48%
Ivanovic, 2005b	Corpus de messages instantanés formé de 550 énoncés.	Anglais	-Lemme -POS -Mot de tête de segment	HMM	WindowDiff = 17%
				Arbre syntaxique	WindowDiff = 13%
Lafferty et al., 2001	Corpus de Penn Treebank constitué de 1,1 millions de mots.	Anglais	Descripteurs orthographiques	CRF	Taux d'erreur = 4,27%
Silvia et al., 2011	Corpus SWITCHBOARD de 1155 dialogues.	Anglais	-Mots -POS -Modèle CoNLL 2000		<i>Accuracy<sub>w</sub></i> = 70,9% <i>Accuracy<sub>T</sub></i> = 58,8%
Elmadany et al., 2015	Corpus JANA formé de 82 dialogues contenant 4725 énoncés	Arabe	-Descripteurs lexicaux -Descripteurs morphologiques -Descripteurs contextuels	SVM	F-mesure = 90,74%
Ngo et al., 2017	Corpus de 3500 messages instantanés	vietnamien	-Descripteurs lexicaux -Descripteurs syntaxiques -POS	CRF	F-mesure=74,51%
				Bi-LSTM-CRF	F-mesure=90,42%
	CRF			F-mesure= 57,27%	
	Bi-LSTM-CRF			F-mesure=73,26%	
(Dalva et al., 2018)	Corpus ICSI-MRDA formé des 73 réunions.	Anglais	-Descripteurs lexicaux -Descripteurs prosodiques	Apprentissage auto-combiné	F-mesure =77,87% Taux d'erreur=42,28%

## 4. Annotation automatique des actes de dialogue

La tâche d'annotation est une tâche primordiale dans le processus de reconnaissance des actes de dialogue. Elle consiste principalement à attribuer une étiquette à chaque énoncé inclus dans les tours de parole des conversations humaines. Nous distinguons deux approches d'annotation d'actes de dialogue à savoir l'approche symbolique et numérique.

Dans cette section, nous présentons les travaux qui s'intègrent dans chacune des deux approches dans le but d'annoter les actes de dialogue dans les discussions Homme-Homme.

### 4.1. Approche symbolique

L'approche à base de règles est fondée sur des règles dégagées à partir de connaissances lexicales (mots clés, phrases clés, etc.), syntaxiques et contextuelles.

(Walker et al., 2001) ont utilisé l'approche à base de règle pour la détection automatique des actes de dialogue. Ils se sont intéressés à la classification des actes dans les dialogues DARPA. Ils ont développé un système basé sur un corpus d'interaction Homme-Machine. Pour se faire, ils ont utilisé la méthode à base de règle RIPPER qui a été appliquée sur les mots des énoncés extraits du corpus annoté avec 11 actes de dialogue. Ils ont obtenu ainsi une valeur d'exactitude égale à 98,5%.

(Prasad et al., 2002) ont repris ce travail, mais en utilisant un corpus Homme-Homme dans le même domaine ; la valeur d'exactitude chute pour 71,8%. Cette diminution est expliquée par le petit changement dans les données introduites au système de reconnaissance automatique. En effet les règles dégagées auparavant ne sont pas compatibles avec le nouveau corpus. Cette défaillance apparaît souvent pour les méthodes d'apprentissage automatique qui sont incapables de supporter la variation des corpus. Cependant (Prasad et al., 2002) ont essayé de trouver une solution à ce problème ; ils ont montré que l'ajout d'un petit ensemble de données (environ 3000 énoncés) pour le corpus utilisé peut augmenter la valeur d'exactitude. Ils ont ainsi dégagé une très bonne valeur d'exactitude égale à 93,8%.

(Lendvai et al., 2003) ont utilisé aussi la méthode RIPPER pour leur corpus OVIS (3738 conversations en Allemand) mais ils ont obtenu une exactitude plus basse de 56,5%.

(Georgila et al., 2009) ont enrichi ce travail par l'ajout manuel de règles spécifiques au contexte pour couvrir le côté utilisateur dans les dialogues. En comparant ce travail avec les données annotées manuellement, ils ont atteint une exactitude d'environ 80%.

## 4.2. Approche numérique

### 4.2.1. Analyse sémantique latente

L'analyse sémantique latente (LSA : *Latent Semantic Analysis*), proposée par (Deerwester *et al.*, 1990), est un modèle algébrique de recherche d'information fondé sur la décomposition en valeurs singulières (*SVD : Singular Value Decomposition*) de la matrice (terme-document) qui représente l'espace d'indexation du modèle vectoriel. Le modèle LSA représente une méthode de récupération de l'information qui a été utilisé pour de nombreux problèmes de classifications.

(Di Eugenio *et al.*, 2014) ont utilisé une extension de la méthode LSA nommée FLSA<sup>1</sup> afin d'ajouter d'autres caractéristiques. Ils ont appliqué cette méthode sur le corpus Spanish CallHome (Levin *et al.* 2003) annoté avec 232 actes de dialogue différents. Pour faciliter la classification automatique, ces actes ont été réduits en 37 et ensuite en 10 actes uniquement. Ils ont utilisé aussi la méthode FLSA avec le corpus MAPTASK (Anderson *et al.*, 1991) qui comprend 13 étiquettes.

(Di Eugenio *et al.*, 2014) se sont basés sur un ensemble de caractéristiques telles que les caractéristiques lexicales, syntaxiques, ainsi que la structure et le contexte de dialogue. En effet, l'enchaînement des actes de dialogue est d'une importance majeure pour la détection automatique de l'acte de dialogue suivant. La méthode LSA a donné un taux d'exactitude de 74,87% pour l'annotation proposée par (Di Eugenio *et al.*, 2014).

### 4.2.2. Modèle N-gram

Le modèles n-gram est l'un des plus simples à utiliser pour la reconnaissance des actes de dialogue. Cette technique est basée sur l'hypothèse que l'acte du dialogue actuel est explicitement déterminé par les k actes de dialogue précédents (Nagata *et Morimoto*, 1994). Par conséquent le n<sup>ième</sup> acte de dialogue candidat est choisi par le principe suivant :

$$C_n = \operatorname{argmax}_c P(C_{n-1}, \dots, C_{n-k+1}) \quad (3)$$

Les probabilités conditionnelles sont extraites du corpus étiqueté en comptant toutes les séquences d'actes de dialogue existantes. Ces probabilités sont calculées par le nombre d'occurrences de la séquence  $(C_{n-k+1}, \dots, C_n)$  dans le corpus d'entraînement divisé par le nombre d'occurrences de la séquence la plus courte.

---

<sup>1</sup>Feature Latent Semantic Analysis

Les valeurs de  $k$  sont souvent égales à 2 ou 3 (c'est pour cela la méthode est parfois nommée aussi bigrammes ou bien trigrammes). Si les dépendances entre les données sont plus connues, les valeurs de  $k$  peuvent être plus grandes.

Le projet de VERBMOBIL (Wahlster, 2000) est le meilleur exemple qui utilise la méthode de n-gram. Ainsi, plusieurs travaux basés sur les modèles de séquence n-grams ont été menés dans le cadre de ce projet à savoir ceux de (Nagata et Morimoto, 1994) et (Reithinger et Klesen, 1997). Ces chercheurs ont obtenu des résultats encourageants ; (Nagata et Morimoto, 1994) ont eu ainsi un score d'exactitude de 77,5% en utilisant les deux meilleures étiquettes (bigrams) et 85,1% en utilisant les trois meilleures étiquettes (trigrams).

(Reithinger et Klesen, 1997) ont appliqué le même modèle que (Nagata et Morimoto, 1994), en utilisant les 18 actes d'annotation de VERBMOBIL avec l'évaluation des n-meilleures approches d'annotation. L'attribution de deux-meilleures étiquettes a donné un taux d'exactitude de 59,62% alors que pour les trois-meilleures étiquettes ils ont atteint une exactitude de 71,93%.

(Core et Allen, 1997) ont étudié les étiquettes de DAMSL appliquées aux dialogues de TRAINS et ils ont conclu que la méthode de séquences d'acte de dialogue diffère selon le contexte utilisé c'est-à-dire elle n'est pas toujours performante.

(Webb et al., 2008) ont utilisé les modèles n-gram pour l'identification des actes de dialogue en se basant sur des caractéristiques lexicales. Ils ont appliqué ces modèles sur le corpus SWITCHBOARD et ils ainsi obtenu une valeur d'exactitude de 54,5%.

### 4.2.3. Chaînes de Markov Cachées

Le processus dans les chaînes de Markov cachées « HMM : Hidden Markov Models » est modélisé comme étant deux séquences d'états parallèles dont l'un est observable et l'autre est caché. Les états de la séquence cachée comprennent une stochastique FSA avec les probabilités de transition spécifiées.

Les deux séquences sont alignées ; les états cachés sont désignés pour générer les états observables, alors que chaque état observable peut être généré par chaque état caché avec une probabilité prédéfinie. Le dernier paramètre, décrivant une instance du modèle, est le vecteur des probabilités initiales des états cachés.

Plusieurs hypothèses sont mises en évidence concernant la nature du processus de dialogue, la première hypothèse apparue est celle de Markov.

La plus forte des hypothèses est que les probabilités de transition et de génération restent constantes tout au long du processus. Les trois paramètres qui décrivent le modèle (la probabilité d'initiation, de transition et de génération) sont estimés à partir du corpus d'entraînement.

L'estimation se fait couramment avec l'algorithme « *forward-backward* » ; l'application du modèle (c'est à dire trouver la plus probable séquence caché étant donné les états observables) est réalisée à l'aide de l'algorithme de Viterbi (Fomey et David, 2005).

L'un des travaux de référence qui ont utilisé les HMMs est le travail de (Stolcke et al., 2000). Ils ont fait une modélisation des actes de dialogue à grande échelle en utilisant le corpus SWITCHBOARD. Les données ont été réparties en deux parties : 1115 conversations comme corpus d'apprentissage et 19 autres comme corpus de test. (Stolcke et al., 2000) ont utilisé une combinaison des chaînes de Markov cachées et les modèles trigrams de séquences d'actes de dialogue. Ils ont obtenu ainsi une exactitude de 71% et 54,3% en éliminant la séquence d'actes de dialogue.

(Surendran et Levow, 2006) ont appliqué aussi l'approche HMM sur les mots de l'énoncé pour la langue Anglaise. Ils ont utilisé le corpus VERBMOBIL, en prenant en compte seulement les données annotées avec les 18 actes de plus haut niveau. Ils ont atteint ainsi une exactitude de 74,7%. Ils ont refait la même expérience sur un corpus allemand, la valeur d'exactitude baisse jusqu'à 67,18%. Cette chute peut être expliquée par la variation de l'ordre des mots dans la langue anglaise par rapport à la langue allemande.

#### 4.2.4. Réseaux bayésiens

L'approche bayésienne est une technique statistique basée sur une forte indépendance à l'hypothèse (Pearl et Russel, 2002). L'idée principale est de maximiser la probabilité de la classe C (la classe d'actes de dialogue) étant donné un ensemble de fonctions  $f_1, \dots, f_n$  représentant l'énoncé:

$$\hat{C} = \operatorname{argmax}_C P(C | f_1, \dots, f_n) \quad (4)$$

Selon le théorème de Bayes, la probabilité à la droite peut être développée comme suit :

$$P(C | f_1, \dots, f_n) = \frac{P(C)P(f_1, \dots, f_n | C)}{P(f_1, \dots, f_n)} \quad (5)$$

Cette approche est souvent utilisée pour des tâches de classification. Plusieurs chercheurs ont exploité les réseaux bayésiens dans la classification et l'annotation automatique des actes de dialogue.

(Grau et al., 2004) ont appliqué une approche bayésienne naïve sur le corpus SWITCHBOARD, ils ont obtenu une valeur d'exactitude de 60% en utilisant le modèle trigram.

(Ji et Bilmes, 2005) et (Ji et al., 2006) ont utilisé les réseaux bayésiens dynamiques (DBNs) avec les modèles graphiques appliqués sur le corpus ICSI-MRDA tout en réduisant le nombre d'actes annotés. Ils ont obtenu des valeurs d'exactitude respectivement égales à 66% et 81% pour leurs travaux de 2005 et 2006. L'amélioration des taux est justifiée par le regroupement des 64 actes utilisés en 2005 en 5 classes uniquement.

(Shala et al., 2010) ont appliqué les réseaux bayésiens naïfs pour l'annotation de discussions humaines avec une taxonomie de 10 actes de dialogue. Ils ont obtenu une valeur d'exactitude de 41,73% pour la classification de 408 segments issus des news collectés du journal arabe Al-Hayat et de la chaîne télévisée Al-Jazeera.

Récemment, (Pleva et al., 2015) ont adopté cette approche en utilisant un corpus d'épisodes télévisés en langue slovaque étiqueté par un schéma d'annotation contenant 12 actes de dialogue. Avec les 18 classes d'actes de dialogue. Ils ont obtenu une valeur d'exactitude de 65% en utilisant des caractéristiques d'ordre lexical, prosodique et contextuel.

### 4.2.5. Réseaux de neurones

Les réseaux de neurones présentent une méthode d'apprentissage supervisé qui a été exploitée dans plusieurs domaines tels que les diagnostics médicaux, la détection des fraudes, ainsi que dans la reconnaissance automatique des actes de dialogue.

Les réseaux de neurones sont composés d'éléments simples (neurones) fonctionnant en parallèle. Ces éléments ont été fortement inspirés du système nerveux biologique.

Comme dans la nature, le fonctionnement du réseau de neurone est fortement influencé par la connexion des éléments entre eux. On peut entraîner un réseau de neurone pour une tâche spécifique (reconnaissance de caractères par exemple) en ajustant les valeurs des connexions (ou poids) entre les éléments.

Les réseaux de neurones constitués de fonctions de transfert « sigmoïdales » et une couche de sortie constituée de fonctions de transfert linéaires sont capables d'approximer n'importe quelle fonction possédant un nombre fini de discontinuité.

La rétro-propagation standard est un algorithme de descente du gradient, comme la loi d'apprentissage de Widrow-Hoff, dans lequel les poids des réseaux sont ajustés dans le sens du gradient négatif de la fonction coût. Le terme de rétro-propagation veut dire que le gradient est calculé pour des réseaux multicouches non linéaires.

(Tran et al., 2017), (Liu et al., 2017) et (Bothe et al., 2018) ont exploité les réseaux de neurones récurrents (RNN) permettant l'annotation des séquences d'actes de dialogue.

En effet, (Tran et al., 2017) ont utilisé les RNNs hiérarchiques en se basant sur les dépendances entre les actes de dialogue. Ils ont ainsi implémenté un réseau à deux couches ayant comme entrée une séquence de mots de l'énoncé à étiqueter et générant en sortie une séquence annotée d'actes de dialogue. Pour se faire, ils ont exploité deux corpus à savoir Maptask (Anderson et al., 1991) et Switchboard (Jurafsky et al., 1998).

Dans première partie, ils ont utilisé 128 conversations du corpus Maptask annotées par un schéma de 13 actes de dialogue présentant des instructions de demande d'information ainsi que des réponses à ces demandes afin de guider l'utilisateur pour atteindre un lieu dans une carte topographique.

Dans une deuxième partie, ils ont exploité 1155 conversations téléphoniques issues du corpus Switchboard avec une taxonomie de 42 actes de dialogue portant sur des thèmes variés.

Ils ont évalué leur système basé sur les réseaux de neurones récurrents, ils ont ainsi obtenu une valeur d'exactitude satisfaisante qui atteint 65,85% pour le corpus Maptask et 75,61% en utilisant le corpus Switchboard. Les résultats ont montré une meilleure performance pour le recours aux RNNs via l'incertitude de propagation.

((Liu et al., 2017) et (Bothe et al., 2018)) ont utilisé aussi les réseaux de neurones pour l'identification des actes de dialogue. Ils ont mis l'accent sur l'aspect contextuel des conversations ; ils ont ainsi intégré des informations portant sur l'historique du dialogue à savoir l'acte précédent et le changement du locuteur.

(Liu et al., 2017) ont testé leur modèle sur une partie du corpus Switchboard (40 conversations : 20 pour l'apprentissage et 20 pour le test). Les résultats obtenus sont encourageants (exactitude= 74,5%) en utilisant uniquement l'apprentissage profond. Ces résultats révèlent une amélioration remarquable lors de l'intégration du contexte d'énonciation (exactitude= 79,7%).

(Bothe et al., 2018) ont évalué les RNNs sur la totalité du corpus Switchboard (1115 conversations pour l'apprentissage et 19 pour le test). Ils ont obtenu une valeur d'exactitude égale à 73,96% sans considération du contexte. Les résultats ont été améliorés lors de l'ajout de caractéristiques contextuelles à savoir l'acte précédent (exactitude = 77,34%).

Ces résultats renforcent l'efficacité de l'intégration de données contextuelles dans le processus de classification des actes de dialogue.

Récemment, ((Martínek et al., 2019) et (Raheja et Tetreault, 2019)) ont exploité le "*deep learning*" pour classifier les actes de dialogue en utilisant les réseaux de neurones récurrents.

(Martínek et al., 2019) ont proposé une méthode permettant la reconnaissance multilingue des actes de dialogue en se basant sur deux topologies standards des réseaux de neurones à savoir les réseaux de neurones convolutionnels (CNN :Convolutional Neural Network) et les mémoires à long court-terme (LSTM: long short-term memory). Ils ont utilisé des critères lexicaux basés sur la distribution des mots dans les dialogues. Ils ont exploité le corpus multilingue utilisé dans le cadre du projet Verbmobil formé de dialogues en anglais et en allemand. Le corpus utilisé est formé de 7425 dialogues anglais contenant environ 11019 actes et de 16125 dialogues en allemand constitués de 33729 actes de dialogue. CE corpus bilingue a été étiqueté avec un schéma contenant 43 actes de dialogues exprimant principalement des demandes, des suggestions et des actes informatifs. L'évaluation du modèle CNN a donné des résultats encourageants (exactitude = 74,4%). Les résultats sont un peu meilleurs en appliquant le modèle Bi-LSTM (exactitude = 74,9%).

(Raheja et Tetreault, 2019) ont aussi appliqué le "*deep learning*" pour l'identification des actes de dialogue. Ils ont intégré des critères d'ordre lexical et contextuel reflétant l'historique du dialogue. La méthode proposée est basée sur les réseaux de neurones récurrents qui ont été testés sur les corpus SWITCHBOARD (Jurafsky et al., 1997) et ICSI-MRDA (Shriberg et al., 2004).

Les résultats d'évaluation sont très encourageants (exactitude=82,9%) pour le corpus SWITCHBOARD et donnent une meilleure valeur d'exactitude (égale à 91,1%) pour le corpus ICSI-MRDA. Ces résultats confirment la performance des réseaux de neurones avec l'intégration du contexte dans le processus de reconnaissance des actes de dialogue.

### 4.2.6. Arbres de décision

Les arbres de décisions sont des approches très populaires pour la classification automatique des actes de dialogue. (Verbree et al., 2006) ont appliqué l'algorithme J48 sur 3 corpus différents : ICSI-MRDA, SWITCHBOARD, et le corpus du projet AMI. Les caractéristiques utilisées étaient les mots transcrits et la ponctuation (par exemple les points d'interrogation). Pour les trois corpus utilisés, le meilleur ensemble de caractéristiques est le même : le point d'interrogation, la présence du mot 'OR' et la longueur de l'énoncé. Avec ces caractéristiques, ils ont obtenu un score d'exactitude de 89,27% pour le corpus ICSI-MRDA, 65,68% pour le corpus SWITCHBOARD et 59,76% pour le corpus AMI.

Récemment, (Blandón et al., 2019) ont utilisé les arbres de décision pour l'extraction des paires question/réponse dans des conversations téléphoniques multilingues (en anglais, allemand et espagnol) contenant 701 questions and 483 réponses. Ils ont exploité un schéma d'annotation formé de cinq types de questions et de sept actes de réponses. Ils ont obtenu une valeur de f-mesure égale à 58% en appliquant des critères lexicaux et sémantiques.

#### 4.2.7. Apprentissage basé transformation

L'apprentissage basé transformation (*TBL :Transformation Based Learning*) est une approche qui a été introduite par (Brill, 1995). Elle est basée sur un ensemble de règles composées de manière supervisée et appliquées consécutivement aux données.

La méthode TBL a été ensuite utilisée par (Samuel et al., 1999) pour la classification automatique des actes de dialogue en utilisant le corpus VERBMOBIL. Ils ont atteint une exactitude de 75,12% tout en appliquant des caractéristiques diverses de l'acte telles que sa longueur, son tour correspondant, etc. Ils ont exploré également l'identification de l'ordre des mots qui pourrait servir comme indicateur clé pour les actes de dialogue.

(Webb et al., 2005)ont repris les caractéristiques utilisées par (Samuel et al., 1999) et ont appliqué la méthode TBL sur le corpus VERBMOBIL en se basant sur les mots clés. Ils ont ainsi obtenu une valeur d'exactitude de 75,12%.

#### 4.2.8. Apprentissage basé mémoire

L'apprentissage basé mémoire (*MBL :Memory Based Learning*) est une méthode d'apprentissage supervisée relativement simple. Pendant la phase d'apprentissage aucun modèle n'est construit, les données sont simplement stockées. Ensuite à la phase de prédiction, la valeur à prédire est déterminée en analysant directement les données déjà mémorisées.

Il existe plusieurs méthodes pour comparer les nouveaux exemples avec ceux déjà stockés, celle la plus fameuse est la méthode de K plus proche voisin (*KNN : k-Nearest-Neighbors*).

Cette méthode vise à mesurer la distance entre les exemples d'apprentissage et les K plus proche voisin exemples déjà classés. Elle assume que les plus "proche" exemples sont similaires au nouvel exemple, qui sera classé ensuite avec les anciens dans la même classe. La façon courante utilisée pour stocker les exemples d'apprentissage est de les remplacer par des fonctions afin de sauvegarder les informations de classification. Pour ce faire, il faut une mesure métrique de distance.

La distance métrique la plus simple pour cette représentation est la mesure « *Overlap* ». Si les caractéristiques individuelles de l'échantillon sont scalaires ou vectorielles, cette mesure utilise

simplement la distance Euclidienne. Si les caractéristiques sont nominales, la distance individuelle est de 1 si les valeurs sont différentes et 0 si elles coïncident.

La combinaison entre le classement k-NN et la mesure « *Overlap* » est appelée l'algorithme IB1. (Lendvai et Geertzen, 2007) ont utilisé IB1 sur un corpus Allemand de 3738 interactions annotées avec 94 actes différents. Cette expérience a utilisé plusieurs caractéristiques en même temps y compris les caractéristiques lexicales, prosodiques et contextuelles. La valeur d'exactitude obtenue est de 73,5%.

(Rotaru, 2002) a appliqué la méthode MBL sur le corpus SWITCHBOARD divisé en deux parties, environ 195000 énoncés pour l'apprentissage et 5000 autres pour le test. Les meilleurs résultats étaient de 72,32% comme exactitude en choisissant les trois plus proches voisins.

Une autre méthode d'apprentissage "paresseuse" appelée TiMBL a été utilisée par (Levin et al., 2003) sur le corpus du projet de traduction automatique NESPOLE. Ils ont utilisé l'algorithme IB1 voisin de 1; Ils ont ainsi obtenu une exactitude de 70%.

### 4.2.9. Séparateurs à vaste marge

Les séparateurs à vaste marge (SVM: Support Vector Machine) sont des techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination. Ces modèles représentent une forme innovante et efficace d'apprentissage à noyau qui ont été appliquées à nombreux domaines à savoir la recherche d'information, la vision par ordinateur ainsi que le domaine de TALN.

(Elmadany et al., 2018) ont appliqué SVM multi-classes pour l'étiquetage de 82 messages instantanés issus de centres d'appels en dialecte égyptien. Ces messages sont collectés à partir des demandes de services auprès des banques, des informations sur les vols aériens ainsi que des services offerts par les opérateurs téléphoniques.

Ils ont intégré des critères d'ordre morphologique tel que la partie du discours, les mots contextuels ainsi que des descripteurs structurels à savoir le type du locuteur (opérateur ou client) ainsi que l'acte précédent qui donnent une idée sur l'acte à prédire. Par exemple, l'acte « acceptation » est généralement suivi d'une demande de confirmation.

Ils ont évalué SVM sur le corpus JANA (ElMadany et al., 2015) formé de 4725 énoncés contenant 20311 mots. Ils ont obtenu des résultats satisfaisants avec la valeur f-mesure de 72,75% pour les messages issus de services bancaires, 64,55% pour les vols aériens et 68,14% pour les opérateurs téléphoniques. Ils ont ainsi eu un score de f-mesure global égal à 70,36%.

Les résultats d'évaluation montrent la fiabilité du classificateur SVM pour la compréhension de dialogues Arabes en dialecte égyptien.

### 4.2.10. Champs aléatoires conditionnels

Les champs aléatoires conditionnels (CRF) sont des modèles probabilistes basés sur une approche conditionnelle permettant d'étiqueter et segmenter les séquences de données.

Plusieurs recherches ont implémenté les modèles CRF pour l'annotation des actes de dialogue vu qu'ils tiennent compte des dépendances entre les séquences de mots des énoncés à étiqueter.

(Kumar et al., 2018) ont exploité ces modèles pour annoter les actes de dialogue. Ils ont utilisé deux corpus différents à savoir:

- Le corpus Switchboard (Jurafsky, 1997) formé de 1155 conversations téléphoniques annotées avec la taxonomie DAMSL(Core and Allen 1997) contenant 42 actes de dialogue.
- Le corpus ICSI-MRDA (Shriberg et al., 2004) qui inclut 75 heures enregistrées à partir de réunions multi-parties annotées avec le schéma MRDA constitué de 39 actes répartis sur 11 catégories générales.

Ils ont intégré des caractéristiques d'ordre lexical en particulier l'étiquette morpho-syntaxique des mots (Part Of Speech: POS) ainsi que des caractéristiques contextuelles reflétant l'enchaînement des actes dans le dialogue afin de prédire les étiquettes correspondantes à chaque énoncé segmenté au préalable. L'intégration du contexte a amélioré les résultats d'évaluation de CRF avec une valeur d'exactitude qui passe de 65,1% à 68,8% pour le corpus Switchboard et une augmentation de 1% via le corpus MRDA (exactitude de 87% à 88%).

(Kumar et al., 2018) ont profité des avantages de l'apprentissage profond en intégrant les réseaux de neurones hiérarchiques Bi-LSTM avec les modèles CRF. Cette combinaison (Bi-LSTM-CRF) a apporté une amélioration de la performance de CRF. Ainsi, la valeur de l'exactitude a atteint 79,2% pour le corpus Switchboard et 90,9% pour le corpus ICSI-MRDA. Ces résultats sont très encourageants et confirment l'efficacité de l'apprentissage profond intégré avec les modèles CRF.

(Chen et al., 2018) ont aussi appliqué les modèles CRF sur les corpus Switchboard et MRDA.

La nouveauté dans ce travail est l'intégration de l'aspect sémantique en plus des descripteurs lexicaux et contextuels. En effet, (Chen et al., 2018) ont incorporé l'inférence sémantique hiérarchique (Attentive Structured Network: ASM) en tenant compte du mécanisme de mémoire dans la modélisation des segments de dialogue. En effet ce mécanisme permet au modèle d'apprendre à plusieurs niveaux de granularité à savoir les niveaux mot, énoncé et

conversation. Ensuite, un réseau interne de dépendances est construit pour encoder l'inférence interne des énoncés et spécifier les dépendances structurelles entre les mots des segments à annoter. La méthode proposée (CRF-ASM) a été implémentée en utilisant les réseaux de neurones récurrents basés mémoire (LSTM) afin de modéliser d'une manière complète les représentations sémantiques des énoncés.

L'évaluation des modèles appliqués a été réalisée par la mesure de la valeur d'exactitude qui a atteint 91,7% pour le corpus MRDA et 81,3% pour le corpus Switchboard. Les résultats obtenus sont très encourageants et se rapprochent bien de l'accord humain ( $\kappa=84\%$ ) pour les conversations téléphoniques de Switchboard.

### 5. Conclusion

La plupart des travaux de recherche sur la reconnaissance automatique des actes de dialogues s'intéressent à l'identification de ces actes pour fournir des informations portant sur des services routiers, ferroviaires, aériens, bancaires ou téléphoniques.

De nombreuses classifications ont été proposées pour l'analyse des dialogues Homme-Homme et Homme-Machine pour différentes langues telles que l'anglais, l'allemand et l'espagnol.

Cependant peu de travaux ont été élaborés pour l'analyse des discussions arabes. Nous citons les travaux de (Shala et al., 2010) et (Elmadany et al., 2015) basés principalement sur des actes de demandes de renseignements et de réponses informatives, confirmatives ou négatives. Ces chercheurs ont testé leurs travaux sur des petits corpus qui sont insuffisants pour juger l'efficacité et la performance de la tâche de reconnaissance d'actes de dialogue.

L'apport principal de notre travail consiste à identifier les actes de dialogue et construire un enchaînement du discours permettant d'extraire les séquences argumentatives à partir des débats politiques arabes. À notre connaissance, il n'y a pas de travaux qui ont traité les actes argumentatifs fréquemment présents dans les débats télévisés en langue arabe.

C'est dans cette perspective que nous mettons l'accent sur les structures argumentatives et communicatives dans les discussions politiques. Nous proposons ainsi un schéma d'annotation spécifique aux débats argumentatifs qui sera par la suite appliqué automatiquement sur un corpus réel d'émissions télévisées.

---

## **CHAPITRE 2**

### **Schémas d'annotation des actes de dialogue**

---

## 1. Introduction

La reconnaissance des actes de dialogue repose sur l'exploitation de corpus généralement annotés par des experts humains dans le cadre de projets d'annotation. Ainsi, la tâche d'annotation se réfère à un ensemble d'actes de dialogue constituant un schéma d'annotation spécifique aux besoins des utilisateurs.

Dans le cadre de notre travail, nous présentons les schémas d'annotation les plus connus ainsi les corpus annotés utilisés dans la plupart des projets d'annotation.

## 2. Aperçu sur les schémas d'annotation et les corpus annotés

### 2.1. Schémas d'annotation

La plupart des taxonomies d'actes de dialogue se sont inspirées de la classification proposée par (Searle, 1968) qu'ils ont adaptée selon leur contexte.

Généralement, ces classifications sont réalisées dans le cadre de grands projets qui s'intègrent dans le contexte de la reconnaissance d'actes de dialogue dans les systèmes de dialogue Homme-Machine ou dans les conversations Homme-Homme. Nous citons les projets MAPTASK (Anderson et al., 1991), VERBMOBIL (Wahlster, 2000), DAMSL (Core et al., 1999), etc.

Dans ce qui suit, nous détaillons les principaux schémas d'annotation proposés afin d'assurer l'annotation automatique des actes de dialogue.

#### 2.1.1. Le projet MapTask

Le corpus utilisé dans le cadre du projet MapTask (Anderson et al., 1991) est formé de 128 dialogues orientés tâche (environ 150000 mots) où une personne vise reconstruire une carte topographique par placement de points de repère en suivant les instructions fournies par une autre personne.

Le projet MapTask utilise 12 actes de dialogue classés en deux grandes catégories:

- *Actes d'initiation*: incluent les instructions, les explications, les questions, etc.
- *Actes de réponse*: regroupent tous les types de réponses, les acquiescements, les clarifications, etc.

Le tableau 2.1 illustre la liste des actes utilisés dans le projet MapTask. Le schéma d'annotation complet de ce projet est détaillé dans le travail de (Carletta et al., 2005). Ces derniers ont réalisé des résultats fiables de point de vue accord entre annotateurs humains. Ils ont obtenu un rapport kappa de 0,83.

<i>Actes d'initiation</i>	<i>Actes de réponse</i>
instruct	acknowledge
explain	reply-y
check	reply-n
Align	reply-w
query-yn	clarify
query-w	ready

*Tableau 2.1: Les actes de dialogue du projet MAPTASK*

### 2.1.2. Le projet VERBMOBIL

Le projet VERBMOBIL (1993-2000) vise à développer un système de traduction automatique de parole pour les langues suivantes: l'allemand, l'anglais américain et le japonais (Wahlster, 2000). Ce projet a été subdivisé en deux phases :

- *Première phase* : annotation d'un corpus de 168 dialogues en anglais (contenant 3117 énoncés) avec les 43 actes de dialogue (Jekat et al., 1995). Ces 43 actes ont été ensuite raffinés et réduits en 18 actes seulement. Le tableau 2.2 présente le détail du nouveau schéma réduit.

<i>Actes de dialogue</i>	
Thank	deliberate
Bye	request-suggest
Greet	request-comment
Suggest	accept
Reject	init
Digress	clarify
give-reason	motivate
garbage	feedback
Confirm	introduce

*Tableau 2.2: Les 18 actes de dialogue de plus haut niveau de VERBMOBIL*

- *Deuxième phase* : élargissement du corpus par l'ajout des dialogues de la planification de plans de vol complets aux dialogues de la planification des réunions. Ce changement de domaine d'application a engendré une nouvelle hiérarchie d'actes de dialogue (Alexandersson et al., 1998). Le nouveau schéma d'annotation comprend 29 étiquettes au lieu des 43 dégagées dans la première phase.

### 2.1.3. Le projet DAMSL

En 1998, les chercheurs de DRI (Discourse Resource Initiative) ont finalisé une collection d'actes de dialogue appelée DAMSL (Dialog Act Mark-up in Several Layers) pouvant être appliquée dans des domaines variés (Core et al., 1999).

Le schéma d'annotation de DAMSL a été utilisé pour l'annotation de plusieurs corpus à savoir TRAINS (Core and Allen, 1997) et SWITCHBOARD (Jurafsky et al., 1998).

L'annotation de DAMSL est multidimensionnelle; elle est décomposée en quatre dimensions:

- *Le niveau informationnel* : inclut les tâches et la gestion de tâches.
- *Le niveau communicatif* : présente les actes qui n'ont pas de relations avec le contenu informationnel du dialogue.
- *La fonction prospective* : inclut les questions, les déclarations, les demandes, les offres, les engagements, etc.
- *La fonction rétrospective* : inclut les réponses, les acceptations, les rejets, etc.

L'accord inter-annotateur, mesuré par le ratio kappa, est calculé pour chaque type d'acte (Allen et Core, 1997). Par exemple :  $k=0,66$  pour l'acte « *déclaration* » ;  $k=0,76$  pour l'acte « *réponse* » et  $k=0,15$  pour l'acte « *engagement* ». Ainsi la fiabilité du schéma d'annotation de DAMSL dépend du type d'acte. Pour cela, il y a peu de dialogues qui ont été annotés avec la hiérarchie complète de DAMSL.

Le détail des dimensions proposées par DAMSL est explicité dans le tableau 2.3.

<p><b>1. Niveau communicatif</b></p> <ul style="list-style-type: none"> <li>• Uninterpretable</li> <li>• Abandoned</li> <li>• Self-talk</li> </ul> <p><b>2. Niveau informationnel</b></p> <ul style="list-style-type: none"> <li>• Task</li> <li>• task-management</li> <li>• Communication-management</li> <li>• Other-level</li> </ul> <p><b>3. Fonction prospective</b></p> <ul style="list-style-type: none"> <li>• Statement <ul style="list-style-type: none"> <li>- Assert</li> <li>- Rassert</li> <li>- other-statement</li> <li>- open-option</li> </ul> </li> <li>• action-directive inuencing-addressee-future-action</li> <li>• info-request</li> <li>• committing-speaker-future-action <ul style="list-style-type: none"> <li>- offer</li> <li>- commit</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>• Conventional <ul style="list-style-type: none"> <li>- Open</li> <li>- Closing</li> </ul> </li> <li>• explicit-performative</li> <li>• exclamation</li> <li>• other-forward-function</li> </ul> <p><b>4. Fonction rétrospective</b></p> <ul style="list-style-type: none"> <li>• agreement <ul style="list-style-type: none"> <li>- accept</li> <li>- accept-part</li> <li>- maybe</li> <li>- reject-part</li> <li>- reject</li> <li>- hold</li> </ul> </li> <li>• understanding <ul style="list-style-type: none"> <li>- signal-non-understanding</li> <li>- signal-understanding <ul style="list-style-type: none"> <li>* acknowledge</li> <li>* repeat-rephrase</li> <li>* completion</li> </ul> </li> <li>- correct-misspeaking</li> </ul> </li> <li>• answer <ul style="list-style-type: none"> <li>information-relation</li> </ul> </li> </ul>
--	--

Tableau 2.3: Les actes de dialogue du projet DAMSL

#### 2.1.4. Le projet AMI

Le projet AMI (Carletta et al., 2005) est un projet de recherche européen centré sur la technologie multimodale des réunions. Le corpus AMI contient 100 heures de réunions capturées à l'aide de nombreux appareils d'enregistrement synchronisés. Il est conçu pour des travaux dans le traitement de discours et de vidéos, ainsi que dans l'ingénierie linguistique, la linguistique des corpus et dans la psychologie organisationnelle. Il a été transcrit avec des sous-ensembles annotés à partir des entités nommées, des actes de dialogue et des gestes multimodaux tels que le simple regard et le mouvement de la tête.

Les actes de dialogue de ce projet se résument en 15 étiquettes détaillées dans le tableau 2.4.

<i>Actes de dialogue</i>	
Back-channel	Assess
Stall	Elicit-assessment
Fragment	Be-positive
Inform	Be-negative
Elicit-inform	Comment-about-understanding
Suggest	Elicit-comment-about-understanding
Offer	other
Elicit-offer-or-suggest	

*Tableau 2.4: Les actes de dialogue du projet AMI*

#### 2.1.5. La taxonomie DIT++

La taxonomie DIT++ (Bunt et al., 2009) est un système complet d'actes de dialogue obtenu par l'extension de la taxonomie de la théorie d'interprétations dynamiques (DIT), initialement développé pour les dialogues d'information, avec un certain nombre d'actes issus du schéma d'annotation du projet DAMSL (Allen et Core, 1997).

L'annotation de DIT++ supporte la multi-dimensionnalité des actes non seulement dans l'assignation de plusieurs étiquettes à un énoncé mais aussi les dimensions qui peuvent représenter plusieurs aspects de la communication adressés indépendamment. Par exemple, chaque énoncé peut avoir plus qu'une fonction communicative (au maximum une fonction par dimension).

Le schéma d'annotation de DIT++ est formé de 11 dimensions avec 95 fonctions communicatives dont autour de 42 sont des fonctions d'ordre général. Le reste des fonctions couvre des actions de gestion de tours de parole ainsi que le contrôle des obligations sociales. Ce schéma a été utilisé dans l'annotation d'une partie des dialogues du projet AMI formé de 57000 énoncés.

Afin de tester la fiabilité de l'aspect multidimensionnel de la taxonomie proposée, (Geertzen et Bunt, 2006) ont annoté une collection de 558 énoncés collectés à partir des corpus OVIS (Lendvai et al., 2003), DIAMOND (Geertzen et al., 2004) et d'une collection de dialogues allemands du projet de (Caspers, 2006).

L'accord inter-annotateurs est mesuré par le rapport kappa inter-dimensions. Ce score montre une fiabilité pour certaines dimensions à savoir la dimension « *gestion de tours* » avec un score de 0,82. D'autres dimensions plus fréquentes dans les corpus présentent un faible accord telle que la dimension « *tâche* » qui ne dépasse pas 0,47. Ces scores montrent que les schémas multidimensionnels peuvent causer divers problèmes d'annotation.

#### 2.1.6. Le Schéma de (Shala et al., 2010)

Il existe peu de travaux qui ont construit des schémas d'annotation pour la langue arabe. Nous citons le travail réalisé par (Shala et al., 2010) qui ont proposé une taxonomie contenant 10 actes de dialogue. Cette taxonomie a été utilisée pour annoter des dialogues pour différents domaines. Le tableau 2.5 illustre la liste des actes proposés par (Shala et al., 2010).

<i>Actes de dialogue</i>	
Assertion	indirect request
Declaration	question
Denial	promise,
expressive evaluation	response to question
Greeting	short response

Tableau 2.5: Les actes de dialogue de (Shala et al., 2010)

#### 2.1.7. Le Schéma de (Elmadany et al., 2014)

(Elmadany et al., 2014) ont proposé un schéma dédié aux renseignements aériens, bancaires ainsi qu'aux informations à partir des messages téléphoniques instantanés.

La taxonomie proposée est répartie en deux principales dimensions à savoir les actes de demande et les actes de réponse.

La dimension « demande » inclut différents types de questions telles que les questions oui/non, les questions de confirmation, les questions de service et les questions de choix.

La dimension « réponse » intègre les acceptations, les rejets, les propositions, les informations ainsi que les actes de gestion des obligations sociales tels que les remerciements et les salutations. Ces actes forment des réponses aux demandes d'information et de services présentées dans la première dimension. Le schéma proposé incluant 25 actes de dialogue est détaillé dans le tableau 2.6.

<i>Actes de demande</i>	<i>Actes de réponse</i>	
Taking-Request	Agree	Correct
Service-Question	Disagree	Pausing
Confirm-Question	Greeting	Suggest
YesNoQuestion	Inform	Promise
Choice-Question	Thanking	Warning
Other-Question	Apology	Service-Answer
Turn-Assign	MissunderstandingSign	Other-Answer
<b><i>Autres actes</i></b>		
Opening		
Closing		
Self-Introduce		

*Tableau 2.6: Les actes de dialogue de (Elmadany et al., 2014)*

## 2.2. Corpus annotés

### 2.2.1. Le corpus SWITCHBOARD

Le corpus SWITCHBOARD est constitué de 200000 énoncés annotés issus de conversations téléphoniques discutant des domaines variés.

(Jurafsky et al., 1998) ont compressé le schéma hiérarchique de DAMSL en une collection d'actes élémentaires atomiques qui regroupent la fonction fournie par l'acte ainsi que l'information hiérarchique issue du schéma initial de DAMSL. Ainsi, la proportion du corpus Switchboard annotée par DAMSL est formée de 1155 conversations contenant 225000 énoncés. Le schéma Switchboard-DAMSL résultant est constitué de 220 actes de dialogue.

Bien que le nombre important d'étiquettes dans la taxonomie proposée peut causer un éventuel désaccord entre annotateurs humains lors de la tâche d'annotation manuelle, (Jurafsky et al., 1998) ont pu réaliser un score moyen de kappa égal à 80%.

Parmi les 220 actes, il y a 130 qui ocurrent moins de 10 fois dans le corpus. Ainsi, les auteurs ont visé à regrouper ces actes en 42 étiquettes exclusives afin d'enrichir le nombre d'occurrences par classe ce qui améliore les résultats d'annotation automatique via des modèles statistiques.

### 2.2.2. Le corpus ICSI-MRDA

Similairement à SWITCHBOARD, le corpus ICSI-MRDA (Shriberg et al., 2004) est annoté en utilisant une variante de la taxonomie DAMSL en combinant quelques actes en une seule catégorie afin de réduire l'aspect multidimensionnel du schéma original.

Le corpus ICSI-MRDA est formé de 75 réunions multi-parties traitant l'ordre du jour des articles discutés d'une manière formelle et structurée. Ce corpus contient environ 180000

énoncés; chaque énoncé est annoté par un acte de dialogue général et zéro ou plusieurs fonctions spécifiques. Ainsi, ce corpus inclut 11 actes généraux et 39 actes spécifiques.

(Ang et al., 2005) ont réduit la multi-dimensionnalité du schéma DAMSL en gardant uniquement 5 dimensions à savoir « *back-channels* », « *place holders* », « *questions* », « *statements* » et « *disruptions* ».

### 2.2.3. Le projet AMITIES

Le projet AMITIES est fondé sur une coopération entre des universités et des compagnies européennes et américaines. Les chercheurs ont collecté un corpus de 1000 dialogues Homme-Homme contenant 24000 énoncés à partir d'appels téléphoniques auprès des banques traitant l'interaction avec les clients en termes de demandes d'informations concernant le solde ou les transactions réalisées ainsi que les déclarations de perte de cartes de crédit.

L'annotation du corpus AMITIES a été réalisée à la base d'une variante du schéma DAMSL appelée XDML (Hardy et al., 2005). Afin de valider la fiabilité de ce corpus, L'accord inter-annotateur kappa est estimé à 0,59 qui est une valeur inférieure à celle du corpus SWITCHBOARD.

### 2.2.4. Le corpus CallHome

Le corpus CallHome (Levin et al., 2003) comprend 120 appels téléphoniques (44628 énoncés et 12066 mots) entre membres de famille et amis. Le schéma d'annotation de départ inclut 232 actes différents dont la plupart présente une faible fréquence dans le corpus.

Comme dans le corpus SWITCHBOARD (Jurafsky et al., 1998), plusieurs catégories ont été fusionnées tels que les actes « déclaration » et « acquiescement » afin d'obtenir un nombre d'occurrences considérable dans le processus d'apprentissage. La compression du nombre d'actes a donné naissance au corpus CallHome37 qui contient uniquement 37 actes où les chercheurs ont gardé plusieurs sous catégories à savoir la catégorisation des questions (oui/non ou rhétoriques).

Dans le corpus CallHome10, ces sous catégories ont été éliminées ce qui a réduit la taille du schéma d'annotation à 10 actes de dialogue dont 8 sont des propres actes (question, réponse, etc.) avec deux actes additionnels « \% » pour les énoncés abandonnés et « \x » pour le bruit.

L'annotation du corpus CallHome est principalement fondée sur les jeux de dialogue du schéma Maptask. Chaque jeu est formé de mouvements dont chacun contient un ou plusieurs actes de dialogue. Ainsi, chaque jeu débute par un mouvement d'initiation qui s'accomplit par une réponse ou un abandon à la fin de la tâche.

### 2.2.5. Le corpus JANA

Le corpus JANA (Elmadany et al., 2014) est formé de 82 dialogues arabes dont 52 sont collectés à partir de dialogues spontanés portant sur des informations des banques et de compagnies aériennes. Le reste des dialogues est formé de messages téléphoniques instantanés. Ce corpus inclut 4725 énoncés qui ont été annotés à la base du schéma d'annotation de (Elmadany et al., 2015) qui regroupe 23 actes répartis principalement en deux catégories: actes de demande et actes de réponse, similairement au schéma Maptask. Les statistiques du corpus JANA sont détaillées dans le tableau 2.7.

Nombre total de dialogues	82
Dialogues oraux issus de banques	26
Dialogues oraux issus de vols aériens	26
Dialogues écrits issus d'opérateurs téléphoniques	30
Nombre total de tours	3001
Nombre total d'énoncés	4725
Nombre total de mots	20113
Nombre de mots par tour	7
Nombre de mots par énoncé	4

Tableau 2.7: Statistiques du corpus JANA

### 2.3. Tableau récapitulatif

Corpus	Langue	Domaine	Nombre de dialogues	Nombre d'énoncés
MAPTASK	Anglais	Informations routières	128	4787
VERBMOBIL	Multilingue	Planning de vols	168	58961
SWITCHBOARD	Anglais	ouvert	1155	223606
ICSI-MRDA	Anglais	Réunions	75	105000
AMITIES	Anglais	Informations bancaires	1000	23412
CallHome	Espagnol	Communications téléphoniques	120	44628
JANA	Arabe	Informations bancaires/vols	82	4725

Tableau 2.8: Tableau récapitulatif des corpus annotés

## 3. Schéma d'annotation SADA

Après l'étude de l'état de l'art des différents schémas d'annotation, nous proposons une taxonomie d'actes de dialogue (**SADA**: **S**cheme **A**nnotation for **D**ebates in **A**rabic) spécifique à la langue arabe et qui met en relief la structure argumentative des discussions (BenDbabis et al., 2012). Ce schéma est inspiré de la taxonomie générale de la langue arabe. Il a été réalisé en deux

versions: une première version complète contenant 40 actes (SADA40) et une deuxième version réduite comportant 19 actes de dialogue.

Nous détaillons dans ce qui suit la taxonomie générale de la langue arabe ainsi que les actes de dialogue du schéma d'annotation proposé.

### 3.1. Taxonomie générale de la langue Arabe

La taxonomie arabe différencie la langue en trois modes de discours qui sont le mode narratif (السرد), le mode descriptif (الوصف) et le mode argumentatif (الحجاج).

Bien que ces trois modes ne soient pas généralement associés, nous nous intéressons plus à dégager l'argumentation dans les discussions spontanées.

En effet, le mode argumentatif est exprimé à l'aide des actes de dialogue (الأفعال الحوارية). Pour la langue arabe, ces actes sont répertoriés en deux catégories qui sont la construction (الإنشاء) et la déclaration (الخير).

En effet, la construction (الإنشاء) comprend généralement des actions qui ne peuvent être jugées ni vraies ni fausses. Cette catégorie est composée de deux sous catégories à savoir la demande (الإنشاء الطلبي) et la non-demande (الإنشاء الغير طلبي).

La demande peut être exprimée à l'aide de l'interrogation (الاستفهام), la négation (النهي), l'ordre (الأمر), l'avertissement (التحذير), etc.

En outre, la non-demande peut être exprimée à l'aide de l'exclamation (التعجب), l'éloge (المدح), la satire (الذم), etc.

La deuxième catégorie, la déclaration (الخير), contient les phrases ou les expressions qui peuvent être vraies ou fausses. En effet une déclaration est une sorte d'assertion d'un fait ou d'un évènement afin de changer les croyances de l'auditeur.

La taxonomie générale de la langue arabe est détaillée dans la figure 2.1.

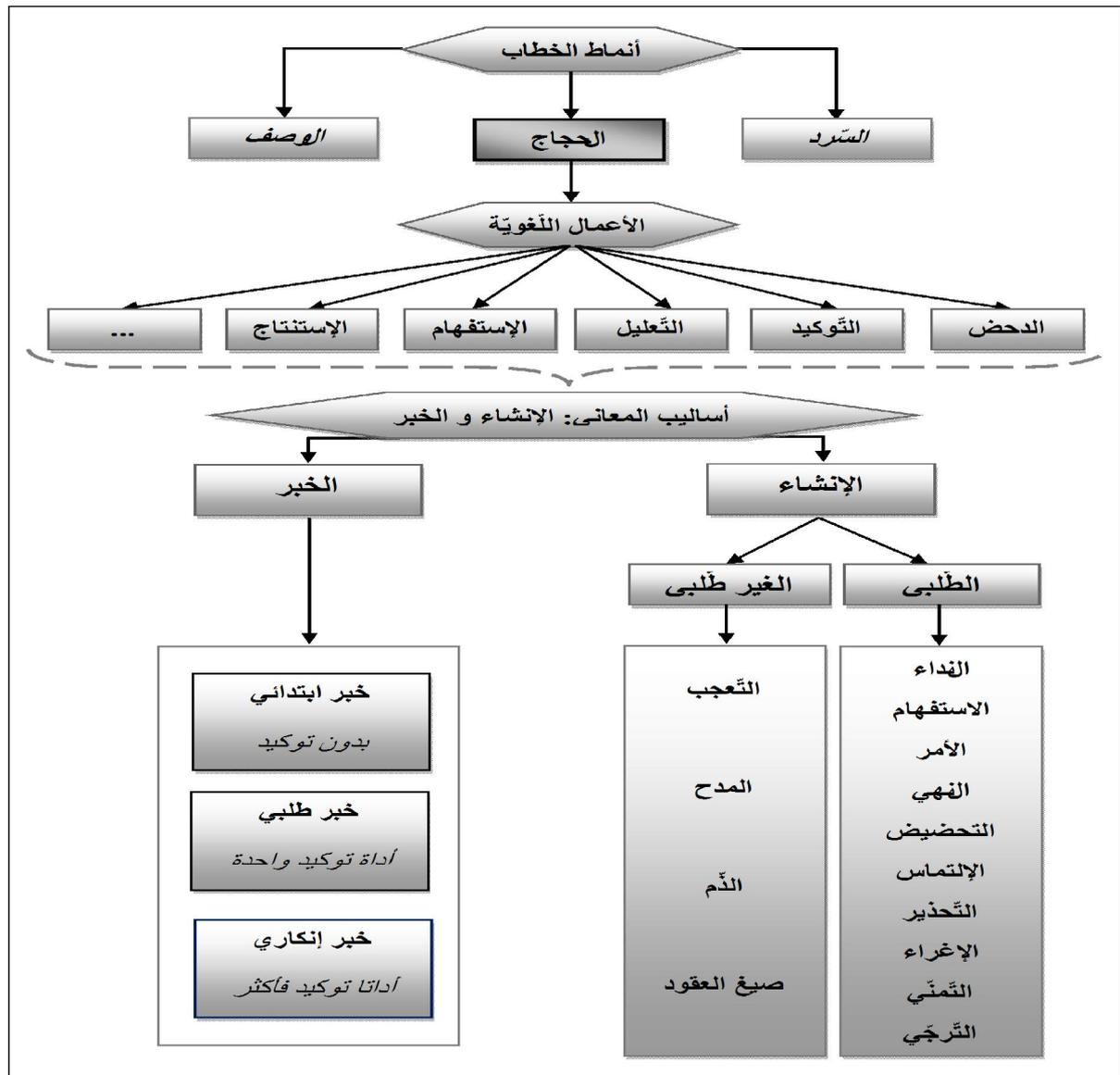


Figure 2.1: Taxonomie générale de la langue arabe

### 3.2. Schéma SADA40

Le schéma SADA40 est formée de 40 actes de dialogue répartis sur six catégories à savoir *la gestion des obligations sociales, la demande, l'argumentation, la gestion de dialogue, la réponse et la déclaration*. Nous détaillons dans ce qui suit les actes associés à chacune de ces catégories.

#### 3.2.1. La gestion des obligations sociales

Cette catégorie contient les actes conventionnels utilisés dans les dialogues tels que les formalités d'ouverture et de clôture, les salutations, ainsi que les actes expressifs proposés par (Searle, 1968) à savoir le remerciement, l'excuse, le regret, etc. Le tableau 2.9 décrit les différents actes de dialogue de cette catégorie.

N°	Actes de dialogue	Description
1	آداب الافتتاح Formalité d'ouverture	Début du dialogue. <ul style="list-style-type: none"> <li>• أهلا بكم مشاهدينا الكرام نحن معكم على الهواء مباشرة في برنامج الاتجاه المعاكس</li> <li>• السلام عليكم ورحمة الله وتحياتي إلى الكل في كل أوطاننا العربية</li> </ul>
2	آداب الإقفال Formalité de fermeture	Clôture du dialogue. <ul style="list-style-type: none"> <li>• مع تقديري لكم، في أمان الله.</li> <li>• وبهذا نكون قد وصلنا إلى نهاية هذه الحلقة من برنامج ما وراء الخبر، إلى اللقاء.</li> </ul>
3	تحية Salutation	Formule de politesse utilisée par l'interlocuteur au début de la parole <ul style="list-style-type: none"> <li>• مساء الخير غسان</li> <li>• شيخ راشد الغنوشي مرحبا بك وأنت في المنفى</li> </ul>
4	تقديم Présentation	Présentation du participant, thème de l'émission, etc. <ul style="list-style-type: none"> <li>• يسعدنا أن نستضيف في هذه الحلقة السيد هيثم يوسف رئيس ملتقى الإعلاميين الشباب العرب.</li> <li>• نايف البشارية معد برامج متخصص في إعداد البرامج الشبابية تلفزيون الراي الكويت.</li> </ul>
5	شكر Remerciement	Expression de remerciement envers quelqu'un. <ul style="list-style-type: none"> <li>• شكرا جزيل لك</li> <li>• أنا شاكر جدا لك</li> </ul>
6	تأسف Regret	Demande de pardon en exprimant un sentiment de regret. <ul style="list-style-type: none"> <li>• للأسف حدود الدول الأوروبية مفتوحة لنا لكن بعض حدود الدول العربية مغلقة أمامنا</li> <li>• كنا نتمنى أن تكون معنا على الهواء مباشرة</li> </ul>
7	اعتذار Excuse	Expression d'excuse pour un mal entente ou une mal expression de l'énoncé précédent. <ul style="list-style-type: none"> <li>• عذرا سيدي</li> <li>• نعتذر مشاهدينا الكرام للتأخير</li> </ul>
8	التماس Demande avec politesse	Demande de quelque chose avec affabilité et d'une manière aimable. <ul style="list-style-type: none"> <li>• أريد أن نناقش هذه النقطة من فضلك</li> <li>• أستسمحكم إذا غفلت عن أي بلد</li> </ul>

Tableau 2.9: La gestion des obligations sociales

### 3.2.2. La demande

Cette catégorie renferme les types de demandes exigeant une réponse auprès du destinataire à savoir la demande de confirmation, le demande d'opinion, la demande de justification, etc. Les différents types de demandes sont explicités dans le tableau 2.10.

N°	Actes de dialogue	Description
9	طلب Demande	
	طلب توكيد Demande de confirmation	Demande d'approuver un énoncé. <ul style="list-style-type: none"> <li>• في العالم كله؟</li> <li>• أليس هذا جميلا؟</li> </ul>
	طلب تفسير Demande d'explication	Demande d'expliquer un énoncé afin d'être plus claire aux auditeurs. <ul style="list-style-type: none"> <li>• كيف تبقى تونس خضرة تونس حرة دون أن يسطو أحد على هبة أو انتفاضة أو ثورة هذا الشعب؟</li> </ul>

	• ما الذي تعنيه الحرية المفرطة؟
طلب تعليل Demande de justification	Demande de donner des raisons, de prouver un énoncé.
	• اليوم أنت كتبت "شكرا للشعب التونسي" ما السبب؟ • لماذا تقول بهذه الطريقة؟
طلب إبداء الرأي Demande d'opinion	Demande d'exprimer un point de vue envers un sujet donné.
	• إذا كانت قضية ساخنة تهمة الأمة قضية العراق قضية فلسطين قضايا أخرى ما الذي يحددها برأيك؟ • ما رأيك سيدي؟

Tableau 2.10: Les types de demandes

Ces demandes peuvent être exprimées à l'aide de différents styles comme l'interrogation, l'impératif, la promesse, l'invocation, etc. Le tableau 2.11 détaille les actes constituant la catégorie demande.

N°	Actes de dialogue	Description
10	الاستفهام Interrogation	Demande d'information du destinataire tout en attendant une réponse. • هل هذا صحيح؟ • ماذا تفعل هذه المليشيات؟
11	الأمر Impératif	Demande directe obligeant le destinataire à faire quelque chose. • جاؤبني على سؤالي • اسكت
12	الوعد Promesse	Engagement de faire quelque chose dans le futur. • الشعوب ستقوم، ستقوم يا رجل • أعدك أن أستضيفه في الحلقة القادمة
13	الترجي Espérance	Attente confiante à la réalisation de quelque chose. • الآن أرجو ألا يفهم الشباب ونحن مع شباب الصحافة والإعلام أن هذه دعوة للتمرد على القوانين • هذا سؤال أرجو أن يوجه للإعلاميين
14	التمني Souhait	Désir que quelque chose s'accomplisse. • أتمنى أن يكون هذا مثل الأيام الوطنية • أتمنى أن يتعظ الزعماء العرب بأن سياسة الحجب و حجب المواقع و تكتيم الأفواه و مصادرة الصحف و محاربة الفضائيات هذه لم تنجح في حماية الدكاتوريين
15	الدعاء Invocation	Prière à Dieu de réaliser un souhait. • وأسأل الله أن يعجل بشفائهم • ندعو الله أن تتحقق كل أهداف الثورة
16	التحذير Avertissement	Appel à l'attention du destinataire à propos de quelque chose • نحن نحذر رموز المعارضة اللذين سيلتحقون بهذه الحكومة بأن لا يكونوا شهداء زور مرة أخرى • توخ الحذر

Tableau 2.11: La demande

### 3.2.3. L'argumentation

L'objectif du discours argumentatif consiste de soutenir un point de vue ou une opinion qui répond à une problématique à propos d'un thème donné. Il faut convaincre l'adversaire, soit pour modifier son opinion ou son jugement, soit pour l'inciter à agir. Pour se faire, l'interlocuteur utilise plusieurs actes de dialogue dans son discours. Le tableau 2.12 illustre les différents actes de dialogue argumentatifs.

Actes de dialogue	Description
17 إبداء الرأى Opinion	Jugement qu'un interlocuteur se forme ou qu'il l'adopte sur un sujet et qui peut être fondé par des arguments.
18 استحسان Appréciation	Jugement favorable envers une opinion ou un énoncé quelconque.
19 استهجان Désapprobation	Jugement défavorable envers une opinion ou un énoncé quelconque.
20 موافقة Acceptation	Réponse affirmative exprimant un accord avec l'énoncé d'un interlocuteur.
21 موافقة جزئية Acceptation partielle	Réponse affirmative exprimant un accord partiel avec l'énoncé d'un interlocuteur.
22 رفض Rejet	Rejet d'une proposition, un jugement ou d'une opinion d'autrui.
23 رفض جزئي Rejet partiel	Rejet d'une partie de la proposition, du jugement ou de l'opinion d'autrui.
24 حجة Argument	Raisonnement tendant à établir une preuve ou à fonder une opinion
25 تعليل Justification	Preuve qu'un interlocuteur donne, comme raison de ce qu'il à dit auparavant.
26 تفسير Explication	Explication d'un point de vue ou d'une proposition.

		<ul style="list-style-type: none"> <li>• يعني لبنان لا يؤخذ بالقوة</li> <li>• أقصد أنه هو هرب كالجبان هرب يعني و هرب لأن الشعب التونسي انتفض</li> </ul>
27	توكيد Confirmation	Deuxième preuve visant à se défendre et à insister sur un point de vue.
		<ul style="list-style-type: none"> <li>• بالفعل منذ بداية عام 2004 بدأ التعاون الفعلي بين الاليزيه والبيت الأبيض</li> <li>• طبعاً نحن همنا الأساسي الآن هم التونسيين والحقيقة هم العرب</li> </ul>
28	استنتاج Conclusion	Conséquence tirée d'un raisonnement proposé par l'interlocuteur.
		<ul style="list-style-type: none"> <li>• إذن الأسد يجب أن يرحل بالنسبة لشيراك</li> <li>• بالتالي الحرية كما استخدم ليست مطلقة هناك معايير للحرية...</li> </ul>

Tableau 2.12: L'argumentation

Pour convaincre l'auditeur, l'interlocuteur doit exposer des idées qui prouvent la justesse de son opinion. Ces idées constituent les arguments qui peuvent prendre plusieurs formes tels que les arguments réels, les exemples, etc. Les différents types d'arguments sont présentés dans le tableau 2.13.

	Types	Description
Arguments	حجة الواقع Argument de la réalité	Argument qui utilise le récit d'évènement pour appuyer une position et illustrer une morale
		<ul style="list-style-type: none"> <li>• وأقرب مثل شهدناه فيما حدث ما بين مصر والجزائر في الرياضة</li> <li>• في السودان في الفترة الماضية كان هناك تشجيع كبير من بعض الصحفيين</li> </ul>
	حجة المماثلة Argument de comparaison	Argument qui établit un parallèle entre deux situations. Il montre leurs points communs pour qu'elles soient considérées et traitées de la même façon.
		<ul style="list-style-type: none"> <li>• الحرية لها ثمن والكرامة لها ثمن</li> <li>• حتى في ثورات كبرى حتى في تاريخ الثورة الفرنسية حتى في إيران عندما حصل قبل ثلاثين عاما أشهر تقريبا لكن شعب تونس استطاع خلال شهر...</li> </ul>
	حجة الشاهد القولي Argument d'autorité	Argument qui s'appuie sur une citation, ou sur une opinion dont on souligne la valeur communément admise.
		<ul style="list-style-type: none"> <li>• في جملة شهيرة الجميع يرددوها، الشعب يريد إسقاط النظام،</li> <li>• كما قال أبو القاسم الشابي إذا الشعب يوما أراد الحياة</li> </ul>
	حجة منطقية Argument basée sur la logique	Argument qui tire son validité de son aspect rationnel et convainc le lecteur par l'adhésion intellectuelle.
		<ul style="list-style-type: none"> <li>• النظرية تقول بأنه ينبغي على الصحفي أن يؤمن بالمثل العليا للصحافة والإعلامية أخلاقيات عمل وعمل مهني</li> <li>• لا يمكن للثورة أن تنجح إلا بتكاتف الشعب و الدولة</li> </ul>
	حجة المثال Argument d'exemple	Témoignage (donnée scientifique, historique, numérique...) destiné à justifier ce que l'interlocuteur dit.
	<ul style="list-style-type: none"> <li>• يعني على سبيل المثال إذا تحدثنا عن بلد معين تجرى انتخابات وهناك من يقول إن هذه الانتخابات ليست نزيهة وإن ثمة تزيف وتزوير</li> <li>• المثال الصارخ عندما حصل ما حصل في الرياضة بين الجزائر ومصر...</li> </ul>	

Tableau 2.13: Les types d'arguments

### 3.2.4. La gestion de dialogue

Cette catégorie renferme les actes de dialogue qui servent à gérer les tours de parole ainsi que l'enchaînement d'un dialogue tels que l'apaisement, l'acquiescement, le signal de non compréhension, etc. Les actes de gestion de dialogue sont détaillés dans le tableau 2.14.

N°	Actes de dialogue	Description
29	قبول Acquiescement	Etre en accord avec l'énoncé précédent
		<ul style="list-style-type: none"> <li>• وصلت الفكرة وصلت الفكرة.</li> <li>• حسنا</li> </ul>
30	علامة عدم فهم Signal de non compréhension	Déclaration exprimant une mal compréhension de l'énoncé précédent
		<ul style="list-style-type: none"> <li>• ماذا؟!!</li> <li>• لم أفهم، لم أفهم ماذا تريد أن تقول!</li> </ul>
31	تهدئة Apaisement	Expression ou geste visant à calmer les interlocuteurs
		<ul style="list-style-type: none"> <li>• لا.. أي مقاطعة أوقف البرنامج،</li> <li>• هدوء من فضلكم</li> </ul>
32	طلب توضيح Demande de clarification	Demande de reformulation d'un énoncé jugé non claire pour l'auditeur
		<ul style="list-style-type: none"> <li>• وضح ليما هي العناوين الأساسية التي يمكن أن تقدموها لتونس الغد؟</li> <li>• ثورة 1952 تقصد؟</li> </ul>
33	توضيح Clarification	Éclaircissement d'un point particulier d'un raisonnement proposé
		<ul style="list-style-type: none"> <li>• بمعنى آخر عندما يقال للصحفي أنت صاحب رسالة وإذا وجدت نفسك في قضية ما يا أخي كن رجلا وتستقبل أخي الحبيب أنا قلت من لديه مشكلة</li> </ul>
34	أسلوب الحكيم Contournement de sujet	Manière de réponse pour s'échapper de donner un point de vue personnel
		<ul style="list-style-type: none"> <li>• لم يصل عندي السؤال</li> <li>• لا أنا كإعلامي لا أتدخل في ما لا شأن لي فيه</li> </ul>
35	تذكير Rappel	Rappeler le destinataire sur ce qui a été évoqué dans les énoncés précédents
		<ul style="list-style-type: none"> <li>• السيدة سهام حذرت من هذا هي اشارت إلى بعض الأحزاب</li> <li>• أنا أشرت في البداية إلى أنك تتابع وتواكب الشأن التونسي على الأقل حسب معلوماتي الشخصية من ثلاثين عاما وربما أكثر من ذلك</li> </ul>
36	حوار مبتور Rupture de parole	Rupture de l'énoncé d'un interlocuteur par une autre personne
		<ul style="list-style-type: none"> <li>• ربما يريد أن يحاكمه وربما يريد أن...</li> <li>• لا هذا يعني...</li> </ul>
37	إرجاء Report	L'interlocuteur reporte un point pour une autre occasion
		<ul style="list-style-type: none"> <li>• سنتطرق إلى هذه النقطة لاحقا</li> <li>• في ما بعد.. في ما بعد...</li> </ul>

Tableau 2.14: La gestion de dialogue

### 3.2.5. La réponse

Suite à une demande, en particulier une interrogation, le destinataire répond. Cette réponse peut être exprimée à l'aide de plusieurs types d'actes de dialogue. Néanmoins, dans certains cas il serait difficile pour l'annotateur de déterminer l'acte de dialogue approprié. C'est pour cette raison que nous avons classé la réponse dans une catégorie à part. Le tableau 2.15 présente des exemples de réponses suite à des questions variées.

N°	Actes de dialogue
38	<p><i>Réponse إجابة</i></p> <ul style="list-style-type: none"> <li>• - متى؟</li> <li>• - في أقرب وقت</li> <li>• - اين؟</li> <li>• - موجودة في الدستور التونسي</li> <li>• - ماذا تقولين في هذه النقطة؟</li> </ul> <p>- غسان نحن كنا مجتمعين طيلة هذا الصباح و طلبنا كجمعيات مدنية هو التنام مجلس تأسيسي يؤسس لدستور جديد</p>

Tableau 2.15: La réponse

### 3.2.6. La déclaration

La déclaration est le discours objectif par lequel on déclare quelque chose. Autrement, c'est une proposition donnée et soutenue comme vraie. En effet, cette catégorie inclut les assertions générales « *non opinion statement* ». Des exemples d'actes déclaratifs sont illustrés dans le tableau 2.16.

N°	Actes de dialogue
39	<p><i>Déclaration خبر</i></p> <ul style="list-style-type: none"> <li>• الرئيس السابق المخلوع الهارب لاذ بالمملكة العربية السعودية</li> <li>• قناة نسمة التونسية تقول مقتل عماد الطرابلسي الشقيق الأصغر لزوجة الرئيس التونسي المخلوع..</li> </ul>

Tableau 2.16: La déclaration

### 3.2.7. Autres

Cette catégorie renferme les actes non interprétables et non classifiables dans la taxonomie proposée. Le tableau 2.17 présente un exemple d'acte inclus dans cette catégorie.

N°	Actes de dialogue
40	<p><i>Autres أخرى</i></p> <p>أنتم يا ناس تونس الخضرة تونس الحرة صنعتم التاريخ بدماء شهدائكم وصبر أحراركم و نِعَم الأخلاق</p>

Tableau 2.17: Autres actes

Dans le but de valider la taxonomie d'actes de dialogue proposée, nous avons mesuré l'accord inter-annotateurs lors de l'annotation manuelle de deux discussions par deux experts.

En effet, L'accord entre les jugements est défini comme la conformité de deux ou plusieurs informations qui se rapportent au même objet. Le taux d'accord ou de « concordance » est estimé par le coefficient Kappa proposé par (Cohen, 1960).

Dans le cas d'une étude d'accord entre deux observateurs statistiquement indépendants ayant  $r$  modalités de jugement, avec  $r \geq 2$ , le coefficient Kappa s'écrit :

$$K = \frac{P_0 - P_e}{1 - P_e} \quad (6)$$

Avec :

- $P_0$ : la proportion d'accord observée.
- $P_e$ : la proportion d'accord aléatoire ou concordance attendue sous l'hypothèse d'indépendance des jugements.

$$P_0 = \sum_{i=1}^r p_{ii} = \frac{1}{n} \sum_{i=1}^r n_{ii} \quad (7)$$

$$P_e = \sum_{i=1}^r p_i \cdot p_{.i} = \frac{1}{n^2} \sum_{i=1}^r n_i \cdot n_{.i} \quad (8)$$

L'accord sera d'autant plus élevé quand la valeur de Kappa est proche de 1 et l'accord maximal est atteint ( $K = 1$ ) lorsque  $P_0 = 1$  et  $P_e = 0,5$ .

(Landis et Koch, 1997) ont proposé un classement de l'accord en fonction de la valeur de Kappa présenté dans le tableau 2.18.

Interprétation	Kappa
Excellent	$\geq 0,71$
Bon	0,70 - 0,61
Modéré	0,60 - 0,41
Médiocre	0,40 - 0,21
Mauvais	0,20 - 0,0
Très mauvais	$< 0,0$

*Tableau 2.18: Degré d'accord et valeur de Kappa*

En effet, pour mesurer le coefficient Kappa nous avons commencé avec une segmentation manuelle et identique pour deux discussions transcrites. Ces deux discussions comportent environ 400 énoncés. Par la suite, nous avons procédé à une annotation manuelle de ces deux discussions à l'aide de deux experts humains. L'accord inter-annotateurs obtenu est de 0,84. Ce score est encourageant et montre une fiabilité de la qualité de la taxonomie proposée. Néanmoins, nous avons raffiné l'évaluation des annotations au sein de chaque classe en mesurant le coefficient kappa par classe. Les résultats obtenus sont détaillés dans le tableau 2.19.

Catégorie	Kappa
Réponse	1
Demande	0,96
Gestion des obligations sociales	0,90
Gestion de dialogue	0,82
Argumentation	0,66
Déclaration	0,47

*Tableau 2.19: Résultats de Kappa par classe*

D'après ces résultats, nous pouvons dire que la plupart des classes sont annotées de façon relativement fiable comme la gestion des obligations sociales et la demande. Néanmoins, nous remarquons une valeur de kappa modérée pour la classe argumentation et une valeur médiocre pour la classe déclaration. Cela peut être jugé par la difficulté et l'ambiguïté qui existent dans l'attribution d'un acte argumentatif à un énoncé donné.

En effet, malgré l'objectivité dans la tâche d'annotation, l'expert humain exprime souvent une subjectivité dans le choix des actes vu la nature de notre corpus d'étude formé de débats télévisés spontanés qui tournent généralement sur des sujets d'actualité.

L'exemple (A) présente une paire d'actes annotée, dont le deuxième acte a été annoté dans un premier annotateur comme opinion et par un deuxième comme explication. Cela revient au degré de compréhension de l'énoncé par l'annotateur et son objectivité dans l'annotation.

(A) Acte1 : « طيب يعني هل أفهم من ذلك الآن عندما تقول بهذه الطريقة إنك الآن تدعو إلى تعديل الدستور أو إلغاء الدستور أو ماذا؟ »

Acte2 : « أنا أدعو لأكثر من ذلك أدعو إلى تفكيك منظومة الاستبداد لا أعتبر أن الذين شاركوا في نظام الاستبداد في الصف الأول »

L'ambiguïté revient généralement au manque d'indices lexicaux qui peuvent dans plusieurs cas aider l'annotateur à décider si par exemple l'interlocuteur est en train d'expliquer objectivement un point quelconque ou bien il est en train de défendre son point de vue.

Bien que l'accord global entre experts sur le schéma proposé est satisfaisant, le processus d'annotation automatique a donné de faibles résultats vu la faible fréquence de certains actes ce qui rend difficile leur reconnaissance par les systèmes d'apprentissage automatique. Ainsi nous avons réduit le schéma initial à 19 actes afin d'améliorer les résultats de détection automatique de ces actes.

### 3.3. Schéma SADA19

Le schéma initial SADA40 a été appliqué sur un corpus de 8 discussions comportant 2365 actes de dialogue. La tâche d'annotation automatique a donné des résultats faibles ce qui nous a mené à élargir le corpus à 22 discussions (6050 actes de dialogue) en un premier temps puis à 50 conversations incluant 14062 actes.

Étant donné que l'objectif principal de l'identification des actes de dialogue est de dégager les structures argumentatives dans les débats politiques, nous avons mis l'accent sur les actes argumentatifs qui sont généralement des réponses à des demandes diverses (demande d'opinion, demande d'explication, question, etc.).

Ainsi, nous avons fusionné les actes exprimant la gestion des obligations sociales (ouverture et fermeture du dialogue, remerciement, salutation, etc.) en un seul acte (SOM). De même, nous avons regroupé les actes utilisés pour la gestion du dialogue en une seule catégorie (TM).

Aussi, nous avons amélioré le schéma initial par de nouveaux actes tel que l'acte « thèse » qui représente une idée généralement énoncée par le présentateur de l'émission qui peut être défendue ou rejetée par les différents participants. Par exemple l'énoncé suivant est annoté par l'acte "thèse/ اطروحة" :

"هناك فراغ أيديولوجي الشباب في السبعينيات كانوا أكثر وعياً سياسياً وأكثر تمرساً في السياسة من شباب اليوم"

Ce type d'acte est généralement suivi par une demande d'opinion afin de renforcer ou rejeter l'idée présentée généralement par l'animateur de l'émission. Ainsi, dans cet exemple nous trouvons l'acte "demande\_opinion/ طلب\_إبداء\_الرأي" à la suite de la thèse énoncée :

هل ترين الصورة بهذه الطريقة؟

Nous avons ajouté l'acte « hypothèse/ فرضية » vu sa présence fréquente lors de l'élargissement du corpus. Cet acte peut être identifié par la présence des mots clés "فرضية", "لو", "فرضا", "إذا".

**Exemples :**

(A1) هذه فرضية من المفروض أنها تكون فرضية تكون خيالية...

(A2) لو تصطدم قناعتني مع وسيلة الإعلام...

(A3) فرضاً أن هذا صحيح...

(A4) إذا كان تعقياً...

Certains actes ont été éliminés puisqu'ils présentent une faible occurrence dans le corpus tel que l'acte « déclaration » que nous avons inclus dans la catégorie « autres ».

Les actes « Appréciation » et « Acceptation partielle » ont été intégrés dans l'acte « Acceptation » vu qu'ils peuvent être considérés comme des formes d'agrément.

Dans le même contexte, nous avons inclus les actes « Dépréciation » et « Rejet partiel » dans la catégorie « Rejet ».

Ces modifications dans le schéma initial ont donné naissance à un nouveau schéma réduit « SADA19 » qui inclut 19 actes détaillés dans le tableau 2.20.

N°	Acte de dialogue	Abréviation
1	Gestion des obligations sociales	Som
2	Gestion des tours de parole	Tm
3	Question	Q
4	Demande de confirmation	Conf_req
5	Demande d'explication	Expl_Req
6	Demande de justification	Justif_Req
7	Demande d'opinion	Op_Req
8	Thèse	th
9	Opinion	Op
10	Acceptation	Acc
11	Rejet	Rej
12	Argument	Arg
13	Justification	Justif
14	Explication	Expl
15	Confirmation	Conf
16	Conclusion	Conc
17	Hypothèse	Hyp
18	Réponse	Ans
19	Autres	oth

*Tableau 2.20: Schéma d'annotation SADA19*

#### 4. Corpus CARD

Le schéma d'annotation « SADA » a été dégagé suite à une étude empirique effectuée sur un corpus de discussions Homme-Homme.

Pour se faire, nous avons opté pour les débats télévisés traitant des sujets politiques d'actualités telles que les révolutions tunisiennes, égyptiennes et libyennes, la guerre en Syrie, etc. Ce choix est justifié par la richesse de ce type de discussions en actes argumentatifs en termes de conflits dans les opinions, d'arguments pour soutenir ou rejeter une thèse proposée, etc.

Ainsi, nous avons collecté une série de discussions issues de la chaîne télévisée « AlJazeera » étant donné que cette chaîne met à disposition les transcriptions textuelles des émissions diffusées. Nous avons nommé ce corpus « CARD » (Corpus of ARabic Debates).

Le corpus utilisé a été acquis en trois étapes donnant naissance à trois versions différentes. Nous détaillons dans ce qui suit les différentes versions du corpus CARD en termes de taille et de schéma d'annotation appliqué.

#### 4.1. Corpus CARD 1.0

En une première étape, nous avons collecté 8 débats politiques que nous avons initialement annotés en utilisant le schéma complet SADA40. Le détail du corpus CARD 1.0 est illustré dans le tableau 2.21.

Nombre de discussions	8
Nombre de tours	773
Nombre d'actes de dialogue	2367
Nombre de mots	37075
Nombre moyen de tours/discussion	97
Nombre moyen d'actes de dialogue/discussion	296
Nombre moyen de mots/discussion	4635
Nombre moyen de mots/acte de dialogue	16
Nombre moyen d'actes de dialogue/tour	4
Nombre moyen de participants/discussion	4

*Tableau 2.21: Statistiques du corpus CARD 1.0*

#### 4.2. Corpus CARD 1.1

En une deuxième étape, nous avons élargi la première version du corpus contenant auparavant 8 discussions.

Nombre de discussions	22
Nombre de tours	1805
Nombre d'actes de dialogue	6050
Nombre de mots	101169
Nombre moyen de tours/discussion	82
Nombre moyen d'actes de dialogue/discussion	275
Nombre moyen de mots/discussion	4599
Nombre moyen de mots/acte de dialogue	16
Nombre moyen d'actes de dialogue/tour	3
Nombre moyen de participants/discussion	6

*Tableau 2.22: Statistiques du corpus CARD 1.1*

Cet enrichissement a donné naissance au corpus CARD 1.1 formé de 22 discussions issues de transcriptions de débats télévisés arabes.

Le corpus CARD 1.1 inclut 6050 actes de dialogue qui ont été annotés via le schéma SADA19. Les statistiques de ce corpus sont illustrées dans le tableau 2.22.

### 4.3. Corpus CARD 1.2

La dernière version du corpus CARD inclut 50 débats politiques contenant 14062 actes de dialogue. Le corpus CARD 1.2 a été annoté avec le schéma SADA19. Le détail de ce corpus est présenté dans le tableau 2.23.

Nombre de discussions	50
Nombre de tours	5085
Nombre d'actes de dialogue	14062
Nombre de mots	260212
Nombre moyen de tours/discussion	102
Nombre moyen d'actes de dialogue/discussion	281
Nombre moyen de mots/discussion	5204
Nombre moyen de mots/acte de dialogue	18
Nombre moyen d'actes de dialogue/tour	3
Nombre moyen de participants/discussion	4

*Tableau 2.23: Statistiques du corpus CARD 1.2*

## 5. Discussion

Plusieurs recherches ont été effectuées pour aboutir à des schémas d'annotation d'actes de dialogue. La plupart de ces travaux proposent des taxonomies d'actes spécifiques à des tâches et domaines donnés. Par exemple, le schéma du projet Verbmobil explicite principalement des actes de planification des vols alors que le projet AMI est centré sur l'annotation multimodale des réunions de travail. En outre, les schémas DAMSL et DIT++, indépendants du domaine, mettent l'accent sur l'aspect multidimensionnel des actes de dialogue.

Ces schémas ont été proposés pour des langues étrangères tels que l'anglais, l'allemand et l'espagnol. Peu de travaux ont traité l'annotation des actes de dialogue pour la langue arabe ; nous citons le schéma proposé par (Shala et al., 2010) basé sur des actes génériques de demandes et de réponses pouvant être appliqués à des domaines variés. Plus récemment, le travail de (Elmadany et al., 2015) est dédié aux renseignements concernant les vols ainsi que les transactions bancaires. Ces schémas implémentés pour la langue arabe proposent principalement des actes de demande d'information suivis de réponses à ces requêtes.

Cependant, dans notre contexte de travail, nous proposons un schéma plus complet avec des actes de dialogue plus complexes afin d'identifier l'aspect argumentatif dans les débats politiques. Ainsi, le schéma que nous proposons repose principalement sur des actes argumentatifs illustrant l'interaction entre les différents participants telles que les idées et les opinions qui peuvent être consolidées par une confirmation et/ou une acceptation ou ignorées par un acte de rejet. D'où l'originalité de notre schéma qui traite les actes argumentatifs dans les débats arabes, riches en conflits entre idées et en arguments pour convaincre l'audience.

## **6. Conclusion**

Nous avons présenté dans ce chapitre un aperçu sur les schémas d'annotation des actes de dialogue ainsi que les corpus annotés via ces schémas. Ensuite, nous avons proposé notre schéma SADA, spécifique à l'annotation des actes argumentatifs fréquents dans les discussions politiques. SADA est constitué principalement de quatre catégories d'actes à savoir la gestion des obligations sociales, la gestion du dialogue, l'argumentation et la demande.

Ce schéma a été testé sur un corpus de transcriptions de débats télévisés arabes que nous avons enrichi en 3 étapes (CARD 1.0, CARD 1.1 et CARD 1.2). La dernière version du corpus inclut 50 discussions contenant environ 14062 actes de dialogue.

Dans le chapitre 3, nous détaillons la méthode proposée pour la reconnaissance automatique des actes de dialogue utilisant le corpus CARD annoté par les étiquettes de la taxonomie SADA.

---

## **CHAPITRE 3**

### **Reconnaissance automatique des actes de dialogue**

---

## 1. Introduction

La reconnaissance des actes de dialogue dans les débats politiques s'avère importante dans diverses tâches telles que l'élaboration des systèmes de Question/Réponse, la réalisation des résumés automatiques des discussions, l'analyse des opinions ainsi que la génération de la structure argumentative reflétant l'enchaînement des actes dans les débats traités.

Le processus de reconnaissance automatique des actes repose sur deux principales tâches à savoir la segmentation en énoncés et l'annotation en actes de dialogue.

Dans ce chapitre, nous proposons une méthode de reconnaissance des actes de dialogue basée sur l'apprentissage supervisé. Ainsi, nous avons appliqué les champs aléatoires conditionnels vu leur efficacité dans les tâches d'étiquetage de données séquentielles. Nous avons exploité le modèle CRF respectivement dans les deux tâches de segmentation et d'annotation automatique des débats télévisés.

Le recours aux modèles d'apprentissage nécessite un grand corpus de discussions segmentées et annotées afin de faciliter la prédiction des nouvelles observations non étiquetées.

Pour se faire, nous avons préparé un corpus favorisant l'apprentissage et l'exploitation des discussions à travers un système d'aide à l'annotation nommé ActAAr: **Act Annotation in Arabic** (BenDbabis et al., 2012). L'annotation des discussions est basée principalement sur le schéma d'annotation SADA (Scheme Annotation for Debates in Arabic) spécifique à la fonction argumentative du dialogue présentée dans la section 3 du chapitre 2 ainsi qu'à la taxonomie générale de la langue arabe présentée dans la section 3.1 du même chapitre.

Nous détaillons ainsi les étapes de la démarche suivie, le corpus d'apprentissage et les caractéristiques utilisées pour la reconnaissance automatique des actes de dialogue.

## 2. Architecture générale du processus de reconnaissance des actes de dialogue

La reconnaissance des actes de dialogue repose principalement sur les étapes de segmentation et d'annotation qui peuvent se faire séparément ou simultanément en une seule étape. Dans le cadre de notre travail, nous avons choisi de suivre une démarche séquentielle (Bendbabis et al., 2018) permettant d'abord la segmentation de la discussion en segments fonctionnels auxquels seront assignés des actes de dialogue dans une étape ultérieure. La figure 3.1 détaille le processus séquentiel de reconnaissance des actes de dialogue.

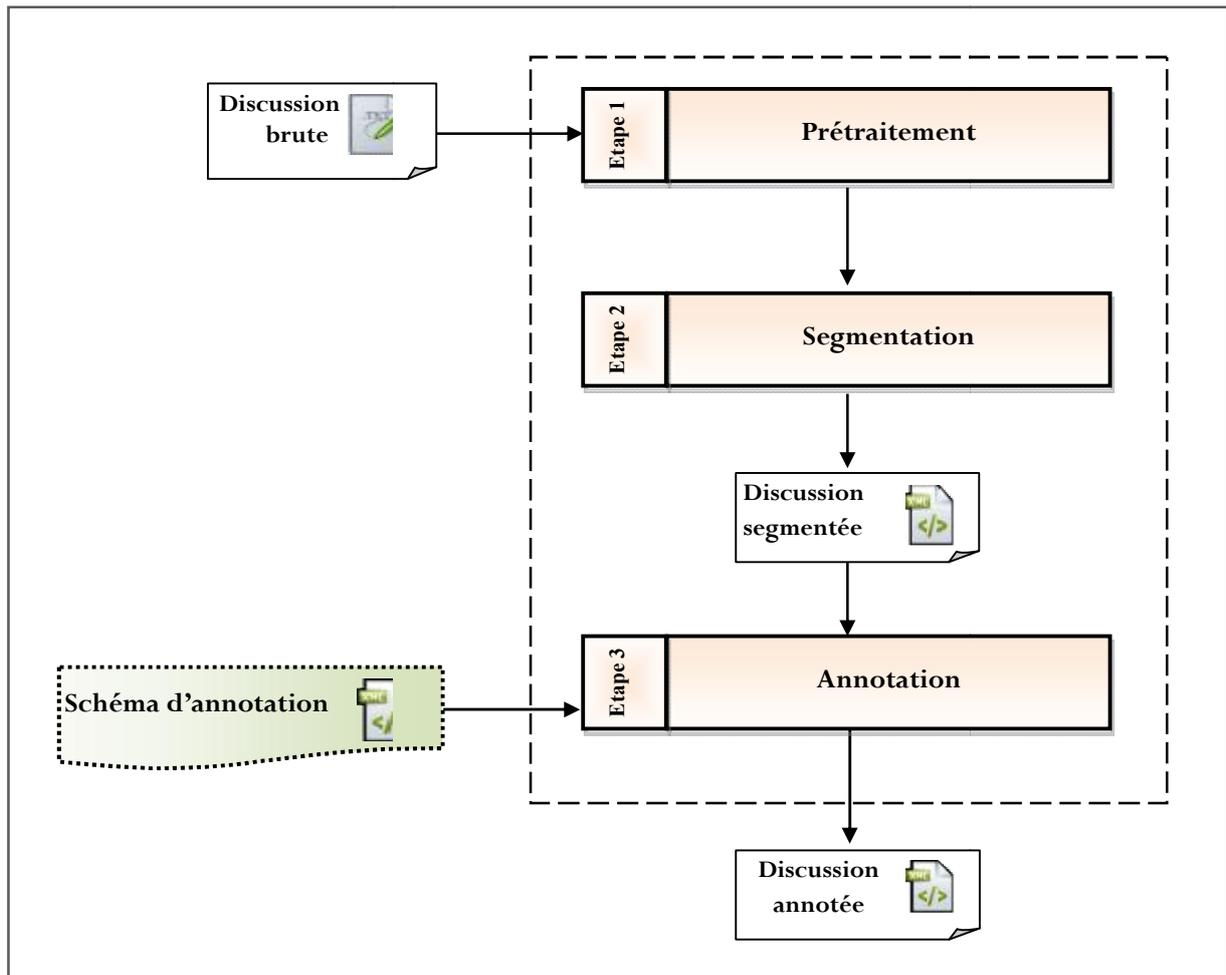


Figure 3.1: Démarche suivie pour la reconnaissance des actes de dialogue

### 3. Apprentissage automatique

L'apprentissage automatique est un processus d'acquisition de savoir-faire qui consiste à développer et implémenter des méthodes permettant d'accomplir des tâches complexes difficiles à réaliser via les moyens algorithmiques classiques.

Ainsi, nous proposons une méthode d'identification des actes de dialogue basée sur des algorithmes d'apprentissage permettant d'apprendre à classifier les actes de dialogue. Cette méthode repose sur l'apprentissage supervisé qui vise à annoter de nouvelles discussions à partir d'un corpus déjà étiqueté.

Pour se faire, nous avons opté pour le choix des champs aléatoires conditionnels (CRF) comme algorithme d'apprentissage pour le processus de reconnaissance automatique des actes de

dialogue. Cette méthode a montré sa performance dans la résolution des problèmes de segmentation, d'annotation et d'étiquetage morpho-syntaxique des données séquentielles.

Plus particulièrement, les modèles CRF ont montré leur efficacité pour la langue arabe. Ils ont été exploités dans la tâche d'analyse syntaxique des textes arabes (Khoufi et al., 2014) ainsi que dans l'étiquetage des dialogues dans le domaine ferroviaire (Graja et al., 2013).

En effet, les modèles CRF offrent plusieurs avantages en termes de vitesse et de taux d'erreur par rapport aux modèles de Markov cachés et les grammaires stochastiques pour ces tâches typiques, y compris la capacité d'atténuer les hypothèses de forte indépendance faites dans ces modèles.

Les champs conditionnels aléatoires sont des modèles graphiques non orientés ayant pour objectif de fournir une distribution de probabilités sur les annotations  $y$  étant donné une observation  $x$ .

Les graphes de HMM, MEMMs (Maximum Entropy Markov Models) et de CRF sont illustrés respectivement dans les figures 3.2, 3.3 et 3.4.

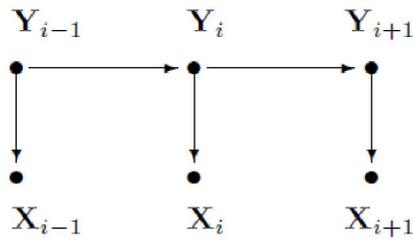


Figure 3.2: modèle graphique de HMM

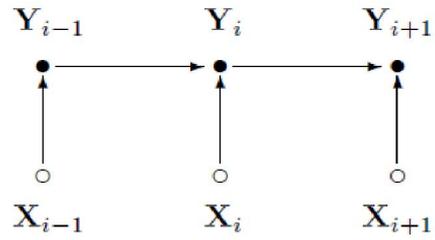


Figure 3.3: Modèle graphique de MEMM

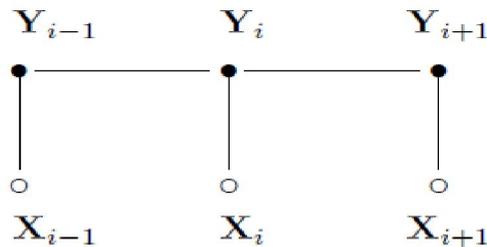


Figure 3.4: Modèle graphique de CRF

(Lafferty et al., 2001) ont définis les champs aléatoires conditionnels comme suit:

Soit  $G = (V, E)$  un graphe non dirigé (graphe d'indépendances) où  $V$  est l'ensemble des nœuds (sommets) et  $E$  l'ensemble des arcs (arêtes).

X et Y étant deux champs aléatoires décrivant respectivement l'observation et son annotation de sorte que pour chaque nœud  $v \in V$  il existe une variable aléatoire  $Y_v$  dans Y.

On dit que  $(X, Y)$  est un champ conditionnel aléatoire si chaque variable aléatoire  $Y_v$  respecte la propriété de Markov suivante :

$$\forall v, p P(Y_v | X, \{Y_w, w \neq v\}) = P(Y_v | X, \{Y_w, (V, W) \in E\}) \quad (9)$$

C'est-à-dire chaque variable aléatoire  $Y_v$  dépend uniquement de X et de ses voisins dans le graphe d'indépendances.

(Lafferty et al., 2001) ont proposé de définir la forme de ces fonctions de potentiel comme l'exponentielle d'une somme pondérée de fonctions  $f_k$  appelées *features*, les  $\lambda_k$  étant les poids associés à chacune de ces *features* :

$$\psi_c(y_c, x) = \exp \left( \sum_k \lambda_k f_k(y_c, x, c) \right) \quad (10)$$

Les *features* sont des fonctions à valeurs réelles. C'est à travers elles que toutes les connaissances du domaine sont intégrées dans le modèle. Dans la plupart des cas, les *features* sont des fonctions binaires valant 1 si un phénomène est observé et 0 sinon. Ces *features* prennent en paramètres les valeurs prises par les variables aléatoires de la clique sur laquelle elles s'appliquent ( $y_c$ ), ainsi que l'ensemble de l'observation x. Par conséquent, la valeur prise par une variable aléatoire peut dépendre de toute l'observation x. Par exemple, dans le cas de l'annotation d'une séquence, le choix de l'étiquette associée au dernier élément de la séquence peut être lié à la valeur du premier élément de cette séquence. À ces *features* sont associés des poids  $\lambda_k$ . Ces poids sont les paramètres du modèle. Ils permettent d'attacher plus ou moins d'importance à certaines *features*, ou même d'indiquer que le phénomène caractérisé par une *feature* ne doit pas se produire (si le poids est négatif).

Un CRF est donc défini par un graphe d'indépendances G et un ensemble de *features*  $f_k$ , auxquels sont associés des poids  $\lambda_k$ . La probabilité conditionnelle d'une annotation connaissant une observation, telle que définie par un CRF, s'exprime alors par :

$$p(y|x) = \frac{1}{z(X=x)} \exp \left( \sum_{c \in C} \sum_k \lambda_k f_k(y_c, x, c) \right) \quad (11)$$

Où Z(x) se réécrit:

$$Z(x) = \sum_y \exp \left( \sum_{c \in C} \sum_k \lambda_k f_k(y_c, x, c) \right) \quad (12)$$

L'intérêt principal de travailler avec un graphe d'indépendances est qu'il permet de mettre en œuvre des techniques de programmation dynamique pour un calcul efficace des deux tâches principales des modèles CRF à savoir la recherche de l'annotation la plus probable et l'apprentissage des paramètres du modèle.

En effet, La recherche de l'annotation la plus probable consiste à trouver l'annotation  $y$  maximisant la probabilité  $P(y|x)$ , étant donnée une observation  $x$  et un CRF dont les paramètres  $\lambda_k$  sont connus.

$$\hat{y} = \arg \max_y p(y|x) = \arg \max_y \sum_{t=2}^T \sum_k \lambda_k f_k(y_{y-1}, y_t, x, t) \quad (13)$$

La tâche d'apprentissage d'un CRF consiste à trouver les poids qui maximisent la log-vraisemblance du modèle, étant donné un ensemble d'apprentissage  $S = \{(x(1), y(1)), \dots, (x(N), y(N))\}$ .

Le modèle CRF peut être appliqué en une seule étape en détectant et étiquetant simultanément les limites des segments de chaque acte comme il peut être appliqué deux fois séparément : segmentation des énoncés puis annotation de chaque énoncé en acte de dialogue.

L'étude expérimentale a montré l'efficacité d'effectuer ces deux tâches séquentiellement l'une après l'autre (F-mesure= 59,2%) au lieu de les regrouper en une seule tâche (F-mesure= 63,4%) vu que les erreurs de segmentation peuvent influencer sur la qualité de l'annotation des actes de dialogue ainsi que certaines caractéristiques d'apprentissage ne peuvent pas s'appliquer sur les deux tâches de segmentation et d'annotation étant donné qu'elles sont pertinentes pour l'une plus que pour l'autre.

Ainsi, nous avons choisi de suivre une approche séquentielle dans la reconnaissance des actes de dialogue. Nous détaillons dans ce qui suit les deux étapes de segmentation et d'annotation des actes de dialogue.

#### 4. Prétraitement

Avant tout traitement de segmentation, il y a des sous tâches préliminaires à appliquer pour que la discussion textuelle soit exploitable et fiable pour des tâches ultérieures telles que l'annotation, l'analyse syntaxique ou sémantique. Le filtrage des discussions textuelles consiste à les transformer en des discussions nettes à travers le traitement de l'entête et des mots ambigus.

#### 4.1. Traitement de l'entête

À chaque débat télévisé transcrit et publié sur le site de la chaîne "AL-JAZEERA "correspond une transcription textuelle. Ainsi, le débat transcrit est constitué de deux éléments à savoir l'entête et le contenu de la discussion.

Vu que la segmentation concerne le contenu du débat, nous avons extrait les informations de l'entête (nom de l'émission, le présentateur, les participants, les sujets discutés, etc.). Un exemple d'entête de discussion est présenté dans la figure 3.5.

<p>اسم البرنامج: حديث الثورة  عنوان الحلقة: عودة رموز النظام السابق بتونس.. أزمة  جديدة بليبيا  مقدم الحلقة: عبد القادر عياض  ضيوف الحلقة:  -نور الدين العلوي/أكاديمي ومحلل سياسي  -صلاح الدين الجورشي/محلل سياسي  -محمود الفطيسي/عضو المجموعة الوطنية الاستشارية  تاريخ الحلقة: 2014/03/09  المحاور:  -فقاعة العودة الإعلامية  -رموز النظام السابق واحتمال عودتهم  -قانون العزل السياسي  -دولة مهلهلة المفاصل</p>
--

Figure 3.5: Exemple d'entête d'une discussion transcrite

#### 4.2. Traduction des mots ambigus

Dans une discussion, chaque participant peut utiliser des mots étrangers à la langue arabe vu qu'il s'agit de l'oral spontané qui peut présenter plusieurs diffuences (par exemple un mot en dialecte tunisien ou un mot en anglais).

Nous avons opté pour la traduction des mots ambigus pour que l'analyse morphologique soit correcte et fiable. Ainsi, cette étape consiste à traduire les expressions empruntées d'autres langues tels que l'anglais ou le français ("ok", "merci", etc.) ainsi que des expressions dialectales provenant du dialecte tunisien ("برشا"/"beaucoup", "متاع"/ à la propriété de, etc.), égyptien ("بس"/"mais", "كويس"/"bon, etc.)ou d'autres dialectes vers la langue Arabe Standard (MSA) comme illustré dans le tableau 3.1.

Expression ambiguë	Expression MSA
يا سيدي الكريم ok	حسنًا يا سيدي الكريم
merci على الاستضافة	شكرا على الاستضافة
No comment أيها السادة	بدون تعليق أيها السادة
لازم يتعلموا و يتوقفوا عن مصادرة الرأي الآخر	يجب أن يتعلموا و يتوقفوا عن مصادرة الرأي الآخر

موش صحيح على خاطر الأمن مهم جدا	غير صحيح لأن الأمن مهم جدا
بس دقيقة و أعطيك الكلمة	فقط دقيقة و أعطيك الكلمة
هايدا الكلام مو حلو	هذا الكلام ليس جميلا

Tableau 3.1: Exemples d'expressions ambiguës traduites en MSA

## 5. Segmentation

La segmentation des discussions Homme-Homme est une tâche clé dans le processus d'analyse automatique du discours. Elle consiste principalement à décomposer la discussion en tours de parole. Chaque tour est par la suite découpé en segments fonctionnels appelés énoncés. À chaque énoncé est assigné un acte de dialogue dans la tâche d'annotation. La figure 3.6 illustre l'enchaînement des étapes de la méthode proposée pour la segmentation automatique des discussions. En une première étape, la discussion est segmentée en tours de parole qui seront par la suite décomposés en énoncés en appliquant l'apprentissage automatique implémentant les modèles CRF.

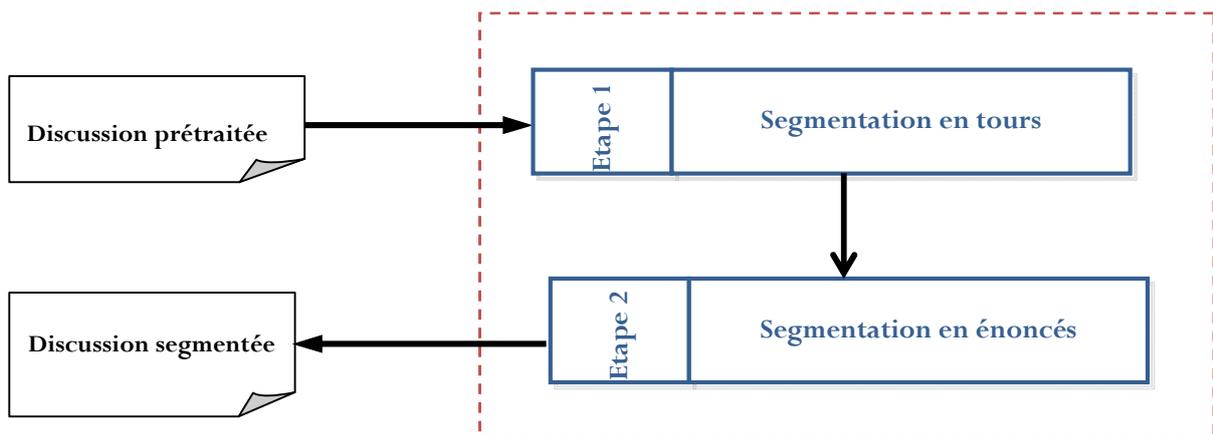


Figure 3.6: Étapes de la méthode de segmentation proposée

### 5.1. Segmentation en tours

La segmentation des discussions en tours de parole est une étape primordiale pour atteindre la segmentation automatique en énoncés puisqu'une discussion est composée d'un ensemble de tours et chaque tour est formé d'un ou plusieurs énoncés.

Chaque tour est énoncé par un interlocuteur qui peut être un participant ou l'animateur de la discussion. Cette segmentation est basée sur la présence du nom de l'interlocuteur suivi par l'indice lexical «:». Ainsi à chaque nouvel interlocuteur correspond un nouveau tour. Un extrait de la segmentation en tours d'une discussion est présenté dans la figure 3.7.

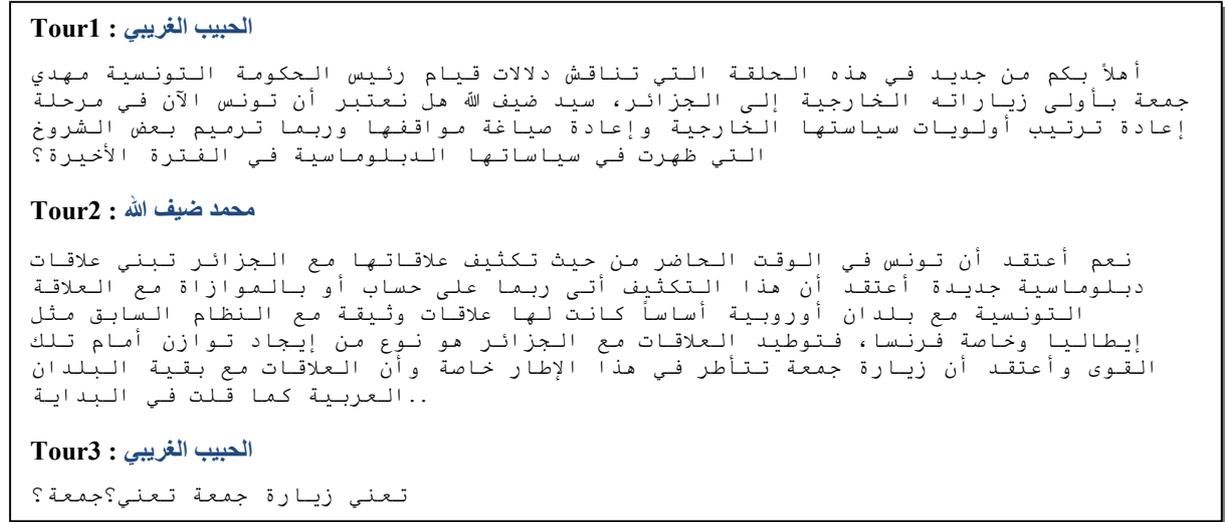


Figure 3.7: Extrait de la segmentation en tours d'une discussion

## 5.2. Segmentation en énoncés

Les discussions segmentées en énoncés facilitent l'annotation automatique de chaque segment de dialogue. Ainsi, l'unité d'énoncé "utterance" est définie par les facteurs suivants :

- Discours prononcé par un seul parlant, sans interruption par un autre interlocuteur, appartenant à un seul tour.
- La complétude syntaxique et/ou sémantique.
- Définit un seul acte de dialogue.

Par conséquent, la disposition d'une base de débats annotés automatiquement facilite la compréhension des opinions des politiciens impliqués dans la discussion en termes de rejets, acceptations, etc.

La segmentation automatique du dialogue en énoncés peut être atteinte à travers l'approche numérique basée sur l'apprentissage ou l'approche symbolique fondée sur des règles.

Bien que cette dernière soit souvent utilisée dans ce contexte, nous avons choisi d'adopter l'approche numérique et plus précisément la technique d'apprentissage supervisé. Ce choix est justifié par le fait que l'apprentissage est doté par sa robustesse et son efficacité en matière de classification des actes de dialogue.

La deuxième raison est que nous disposons d'un corpus d'apprentissage important grâce aux transcriptions textuelles disponibles au site d'Al-Jazeera et à l'outil ACTAAr (bendbabis et al., 2012) qui permet la segmentation semi-automatique des discussions publiées dans ce site.

Nous avons choisi les champs aléatoires conditionnels comme méthode d'apprentissage supervisé permettant l'identification des limites de segments. Ainsi le processus de segmentation automatique est subdivisé en deux phases sous-jacentes à savoir la phase d'apprentissage et celle d'utilisation. De ce fait, nous avons subdivisé le corpus collecté en deux parties:

- Un corpus d'entraînement qui représente presque 80% du corpus total soit 208307 mots qui va servir pour apprendre le système pour produire un modèle de classification.
- Un corpus de test qui représente environ 20% soit 52076 mots pour segmenter de nouvelles discussions à base du modèle d'apprentissage généré par la phase d'apprentissage.

### 5.2.1. Phase d'apprentissage

La phase d'apprentissage automatique nécessite la préparation d'un corpus d'entraînement et d'une base de critères d'extraction (descripteurs) pour pouvoir entraîner le classifieur CRF utilisé pour segmenter les discussions en énoncés.

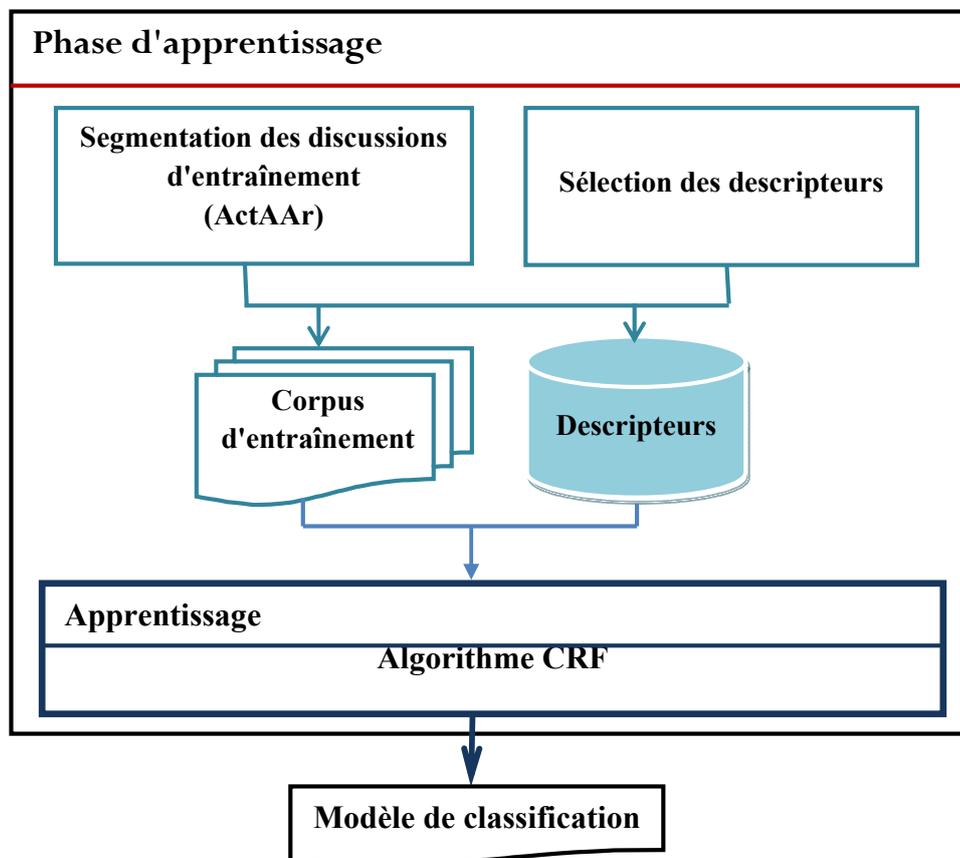


Figure 3.8: Étapes de la phase d'apprentissage

Ainsi, la phase d'apprentissage est composée de 3 étapes: la préparation de corpus d'entraînement, la sélection des descripteurs et l'apprentissage des données. Le détail de ces étapes est présenté dans la figure 3.8.

#### **5.2.1.1. Préparation du corpus**

Chaque système de segmentation ou de classification des actes de dialogue utilise un corpus d'entraînement étiqueté pour accomplir une tâche spécifique à savoir le corpus *SWITCHBOARD* (Jurafsky et al., 1998) contenant 1155 conversations téléphoniques et le corpus *ICSI-MRDA* (Guz et al., 2010) formé de 75 réunions de travail. Ces corpus se réfèrent à des schémas d'annotation spécifiques au contexte de chaque projet.

Vu le manque de corpus étiquetés pour la langue arabe, nous visons préparer un corpus d'étude qui s'applique à notre contexte de travail. Dans cette perspective nous avons choisi de travailler sur les programmes télévisés et plus précisément sur les débats transcrits présentés par la chaîne « Al-Jazeera ». Ces discussions sont disponibles sous forme d'enregistrements vidéo et de transcriptions textuelles disponibles sur le site officiel de la chaîne.

Ce choix se justifie par la disponibilité d'un grand nombre de transcriptions dans le site ainsi que la richesse du corpus collecté en informations argumentatives facilitant l'analyse de la structure communicative de ces discussions.

La taille totale du corpus collecté est de 50 discussions transcrites contenant 260383 mots que nous avons subdivisées en deux parties: une première partie pour l'entraînement des données formant 80% du corpus total (environ 42 discussions contenant 208 307 mots) et une deuxième partie pour l'exploitation et le test du modèle d'apprentissage présentant 20% de la totalité des discussions (soit 8 discussions non segmentées composées de 52 076 mots).

Le corpus d'entraînement a été préparé à travers une segmentation semi-automatique via l'outil ACTAAr qui effectue le marquage des limites de chaque segment d'énoncé pour garantir un fonctionnement efficace de l'étape d'apprentissage implémenté par les modèles CRF.

#### **5.2.1.2. Sélection des descripteurs**

Après la définition des limites des énoncés, nous avons procédé à la sélection des descripteurs d'ordre lexical et morpho-syntaxique à savoir les mots clés, les mots de question, les signes de ponctuation et la partie du discours (POS). Les caractéristiques proposées sont soit de type numérique ou plus exactement binaire telles que le mot clé, le mot de question et la ponctuation soit de type chaîne de caractère comme le critère POS.

- *Mots Clés*

Les mots clés présentent le critère le plus pertinent permettant de délimiter les segments de chaque tour de parole. En effet, ces mots ocurrent fréquemment au début des énoncés pour exprimer une opinion ("أعتقد"/je pense), un accord ("نعم"/oui, "أوافق"/je suis d'accord), un rejet ("لا"/non, "أعترض"/je rejette) ou pour présenter une salutation ("مرحبا"/salut, "تحياتي"/mes salutations), un remerciement ("شكرا"/merci) ou encore pour gérer l'enchaînement du débat("طيب"/bien, "فاصل"/publicité, "هدوء"/silence). La liste des mots clés est lemmatisée pour éviter la conjugaison et la dérivation des mots. Ainsi le lemme de chaque mot est comparé au lemme du mot clé. La valeur de ce critère est binaire: 1 si le mot courant est un mot clé et 0 sinon. Un extrait des mots clés est présenté dans la figure 3.9.

```
<cuewords>
<cueword id="1">رأى</cueword>
<cueword id="2">أعتقد</cueword>
<cueword id="3">أعتبر</cueword>
<cueword id="9">موقف</cueword>
<cueword id="9">رأى</cueword>
<cueword id="13">شكر</cueword>
<cueword id="14">جميل</cueword>
<cueword id="15">طيب</cueword>
<cueword id="17">نعم</cueword>
<cueword id="16">لا</cueword>
<cueword id="18">حسن</cueword>
</cuewords>
```

Figure 3.9:Extrait des mots clés

- *Mots de question*

Les mots de question introduisent souvent une interrogation d'ordre général ou une demande spécifique telles que les demande d'explication, de justification, de confirmation, d'opinion, etc. Ainsi la présence de ces indices lexicaux est souvent indicatrice du commencement d'une question ou demande. Donc nous avons adopté ce critère surtout pour identifier les limites de tout type de demande et de question. Un extrait des mots de question est présenté dans la figure 3.10.

```

<WH_words>
<WH_wordid = "1">ما هي</WH_word>
<WH_wordid = "2">لماذا</WH_word>
<WH_wordid = "3">ماذا</WH_word>
<WH_wordid = "4">ما هو</WH_word>
<WH_wordid = "5">عما</WH_word>
<WH_wordid = "6">ما</WH_word>
<WH_wordid = "7">أين</WH_word>
<WH_wordid = "8">كيف</WH_word>
<WH_wordid = "9">هل</WH_word>
</WH_words>

```

*Figure 3.10: Mots de question*

- *Signes de ponctuation*

Les signes de ponctuation sont importants pour identifier la fin d'un segment de dialogue. D'après une analyse du corpus d'étude segmenté, nous avons remarqué que la plupart des énoncés se terminent par un signe de ponctuation qui peut être un simple point pour clôturer une idée, un argument ou une opinion ou un point d'interrogation pour finir un énoncé interrogatif. Ainsi, le mot qui suit une ponctuation représente souvent le mot de début d'un nouvel énoncé. Le critère de ponctuation prend la valeur 1 si le mot courant est un signe de ponctuation et 0 sinon. Nous présentons dans la figure 3.11 la liste des signes de ponctuation trouvés dans les discussions transcrites.

```

<punctuations>
<punc id="1">./</punc>
<punc id="2">?</punc>
<punc id="3">.</punc>
<punc id="4">,</punc>
<punc id="5">?</punc>
<punc id="6">!</punc>
<punc id="7">;</punc>
<punc id="8">-</punc>
<punc id="8">:</punc>
</punctuations>

```

*Figure 3.11: Liste des signes de ponctuation*

- *POS*

L'étiquetage morpho-syntaxique (POS: Part Of Speech) favorise l'identification des limites des énoncés dans les discussions politiques vu qu'il fournit des informations étroitement liées à la structure syntaxique d'une phrase ou d'un segment de dialogue (Stolcke and Shriberg, 1996). En effet, dans la langue arabe, certaines catégories grammaticales sont reconnues par leur présence au début des segments tels que les pronoms démonstratifs ("هنا"/ce, "هذه"/cette), les verbes

("أشكر"/je remercie, "أعتقد"/je pense), les interjections ("نعم"/oui) ainsi que quelques conjonctions de subordination ("إن", "لقد", "لكن"/mais).

La figure 3.12 présente un extrait de la liste des étiquettes utilisés pour réaliser l'analyse morphologique d'une discussion.

```
<posTag id= "1" >noun</posTag>
<posTag id= "2" >adj</posTag>
<posTag id= "3" >adv</posTag>
<posTag id= "9" >adv_interrog</posTag>
<posTag id= "4" >verb</posTag>
<posTag id= "5" >part_interrog</posTag>
<posTag id= "6" >part_neg</posTag>
<posTag id= "7" >prep</posTag>
<posTag id= "8" >abbrev</posTag>
<posTag id= "9" >punc</posTag>
<posTag id= "10" >conj</posTag>
<posTag id= "11" >conj_sub</posTag>
<posTag id= "12" >interj</posTag>
<posTag id= "13" >pron_interrog</posTag>
```

Figure 3.12: Extrait de la liste des POS

### 5.2.1.3. Apprentissage

Le module d'apprentissage permet la génération d'un modèle d'apprentissage qui prend comme entrée le corpus de débats segmenté en énoncés ainsi que la liste des descripteurs indicateurs des limites de segments de dialogue. Pour se faire, nous avons appliqué CRF comme méthode d'apprentissage supervisé. Le choix de ce classifieur se justifie par sa robustesse en matière de vitesse d'exécution et de classification binaire. En effet, chaque mot de la discussion est étiqueté par l'une des classes suivantes:

- Classe *B (Begin)* qui indique qu'un mot est présent au début d'un énoncé.
- Classe *I (Inside)* qui présente un mot existant à l'intérieur d'un segment de dialogue.

La tâche d'apprentissage d'un CRF consiste à trouver les poids qui maximisent la log-vraisemblance du modèle. Ainsi le problème de classification revient à prédire les étiquettes (B/I) des mots formant l'énoncé à segmenter à partir des caractéristiques d'apprentissage (*features*) d'ordre lexical et morpho-syntaxique.

### 5.2.2. Phase de test

La deuxième phase de l'apprentissage supervisé consiste à exploiter le modèle de prédiction fourni par l'étape d'apprentissage pour segmenter de nouvelles discussions brutes. Ainsi, nous avons utilisé 20% du corpus collecté de débats soit 52076 mots pour la validation du modèle d'apprentissage. La figure 3.13 illustre la phase de test qui se base sur le modèle de classification.

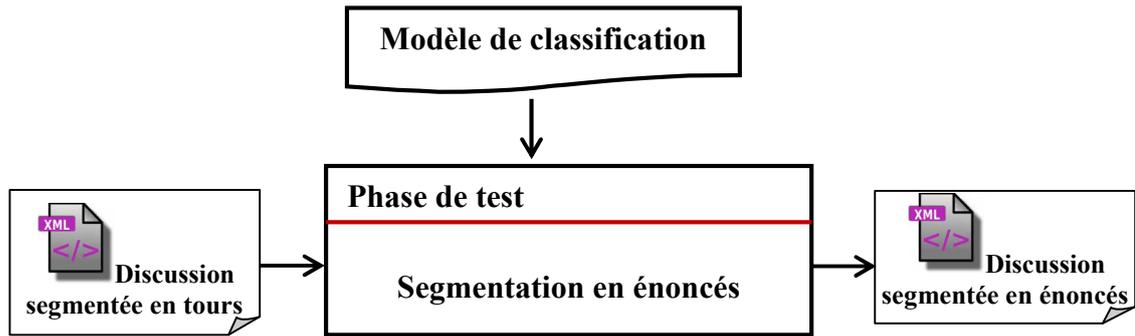


Figure 3.13: Phase de test

Le résultat de cette phase est une discussion segmentée en tours et en énoncés. Un extrait d'une discussion segmentée est présenté dans la figure 3.14.

```

<Turn ID="79" Speaker="غسان بن جدو">
<UtteranceID="50">أين؟</Utterance>
</Turn>
<Turn ID="80" Speaker="راشد الغنوشي">
<UtteranceID="51">موجودة في الدستور</Utterance>
</Turn>
<Turn ID="81" Speaker="غسان بن جدو">
<UtteranceID="52">طيب</Utterance>
<UtteranceID="53">عفوا شيخ راشد</Utterance>
<UtteranceID="54">يعني هل أفهم من ذلك الآن عندما تقول بهذه الطريقة إنك الآن تدعوا إلى تعديل الدستور أو إلغاء
    الدستور أو ماذا؟</Utterance>
</Turn>
    
```

Figure 3.14: Extrait d'une discussion segmentée en énoncés

## 6. Annotation des actes de dialogue

Les corpus annotés sont bénéfiques pour diverses tâches tel que l'établissement des statistiques calculant la fréquence de chaque type d'acte de dialogue. Ces corpus sont souvent exploités comme entrées des modèles d'apprentissage afin de faciliter la reconnaissance automatique des actes de dialogue.

Dans le contexte de notre travail, nous nous intéressons à la tâche d'annotation automatique des actes de dialogue à travers l'utilisation d'un large corpus annoté.

L'étape d'annotation consiste à assigner à chaque énoncé un acte de dialogue. En effet chaque énoncé sera enrichi par des informations supplémentaires indiquant l'acte de dialogue approprié et sa position dans la discussion. La liste des actes de dialogue à assigner est extraite à partir du

schéma d'annotation utilisé qui peut correspondre soit au schéma complet SADA40 formé de 40 actes, soit au schéma réduit SADA19 incluant uniquement 19 actes de dialogue.

Par exemple l'énoncé "مرحبا بكم مشاهديننا الكرام" / *Bienvenue chers spectateurs* sera annoté par l'acte "Remerciement" dans le schéma SADA40 et par l'acte générique "SOM" qui exprime la gestion des obligations sociales d'une manière générale dans le schéma réduit SADA19.

Pour assurer le processus d'annotation automatique, nous avons opté pour les modèles CRF vu qu'ils ont montré leur efficacité dans la tâche de segmentation des discussions en énoncés. Ce modèle est souvent utilisé pour l'étiquetage morpho-syntaxique des textes qui revient à prédire le POS d'un mot dans une phrase.

Pour se faire, nous avons préparé un corpus de 50 discussions contenant environ 14000 actes découpé en deux parties: une première partie pour l'apprentissage et une deuxième pour le test.

### 6.1. Phase d'apprentissage

L'étape d'apprentissage consiste à générer un modèle à partir de la liste des caractéristiques d'apprentissage et des actes de dialogue annotés du corpus d'entraînement. Pour générer ce modèle d'apprentissage, un ensemble de règles doit être établi prenant en considération les dépendances entre les actes annotés pour chaque critère d'apprentissage. La meilleure combinaison de règles donne les meilleurs résultats. Le processus d'apprentissage automatique est détaillé dans la figure 3.15.

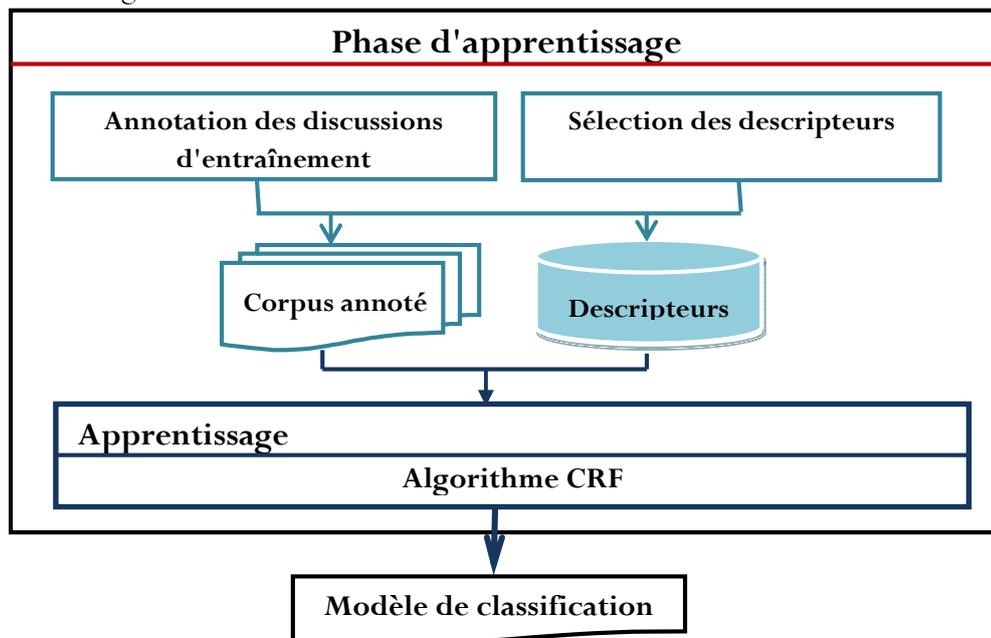


Figure 3.15: Phase d'apprentissage

### 6.1.1. Préparation du corpus

Afin de collecter un corpus d'apprentissage important, nous avons opté pour les débats télévisés traitant des sujets politiques d'actualité telles que les révolutions tunisiennes, égyptiennes et libyennes, la guerre en Syrie, etc. Ce choix est justifié par la richesse de ce type de discussions en actes argumentatifs en termes de conflits dans les opinions, d'arguments pour soutenir ou rejeter une thèse proposée, etc.

Ainsi, nous avons annoté une série de discussions issues de la chaîne télévisée « Al-Jazeera » étant donné qu'elle met à disposition les transcriptions textuelles des émissions diffusées. Nous avons nommé ce corpus « CARD » (Corpus of **AR**abic **D**ebates). Le corpus utilisé a été acquis en trois étapes donnant naissance à trois versions différentes déjà détaillées dans le chapitre précédent.

### 6.1.2. Sélection des descripteurs

Le choix des caractéristiques d'apprentissage est très important dans l'évaluation de la performance des classificateurs utilisés pour la prédiction des données.

La plupart des travaux d'annotation automatique utilisent des descripteurs linguistiques à savoir les critères lexicaux, morphologiques, syntaxiques et sémantiques.

Dans notre contexte de travail, nous avons fait recours à des critères lexicaux tels que les mots clés, des critères morpho-syntaxiques telle que la catégorie grammaticale des mots (POS), des critères niveau énoncé à savoir le rôle du participant et le locuteur de l'énoncé en cours. Nous avons aussi intégré des critères qui reflètent le contexte de la discussion tel que l'acte précédent.

#### 6.1.2.1. Descripteurs lexicaux

Les indices lexicaux sont des critères discriminants pour l'identification des actes de dialogue. En effet, leur présence fréquente dans les segments de la discussion peut donner une idée sur l'acte à prédire. Les descripteurs lexicaux les plus adaptés sont les mots clés qui ocurrent fréquemment dans les discussions.

Par exemple les mots « مرحبا/bienvenue », « شكرا/merci », « إلى اللقاء/au revoir » sont souvent utilisés pour la gestion des obligations sociales.

Les expressions « من فضلك/s'il vous plait », « فاصل و نعود/publicité et on revient », « إهدأ/calmez-vous » et « حسنا/d'accord » sont employées pour la gestion des tours de parole entre participants.

Les mots de question sont considérés aussi comme des indices lexicaux déterminant les différents types de demande telle que la demande de justification débutant souvent par le mot clé « لماذا /pourquoi » ainsi que la demande d'explication utilisant le mot clé « كيف /comment ».

Les mots d'opinion sont aussi utilisés comme des descripteurs lexicaux surtout dans les séquences argumentatives afin d'exprimer une opinion, une acceptation, un rejet ou un argument.

En effet, les expressions « أعتقد /je pense », « برأبي /à mon avis » sont exploitées pour énoncer un point de vue sur une idée donnée.

Les indices lexicaux « أشاطر الرأي /je partage l'avis », « أجزم /j'affirme » sont utilisés généralement pour affirmer une proposition. Le rejet d'une idée est souvent présenté par les mots de négation « لا أوافق /je ne suis pas d'accord ».

Les descripteurs lexicaux ocurrent souvent au début des actes de dialogue mais peuvent être présents à l'intérieur de ces actes.

**Exemples :**

<p>أيا كان الأمر <b>نعتقد</b> أنه لا إمكانية للعودة إلى الوراء.          Quel que soit le cas, <b>nous croyons</b> qu'il n'y a pas de possibilité de revenir en arrière.</p>
<p>سيد رياض أتوجه هنا بالسؤال لك <b>لماذا</b> النقاش في تونس يرتكز على كيفية إدارة الدولة من الناحية السياسية؟          Monsieur Riad, je vous demande <b>pourquoi</b> le débat en Tunisie est basé sur la gestion politique de l'Etat?</p>

Tableau 3.2: Exemples de mots clés présents à l'intérieur des actes de dialogue

### 6.1.2.2. Descripteurs morpho-syntaxiques

La catégorie grammaticale des mots telle que la partie du discours (POS) donne une idée sur l'étiquette de l'acte de dialogue.

En effet, les étiquettes associées aux mots de question renforcent la reconnaissance automatique des actes de questions et des différents types de demandes. Ainsi les demandes sont souvent exprimées par des adverbes interrogatifs tels que لماذا وكيف. Les pronoms interrogatifs هل et ماذا indiquent généralement des actes de questions.

Certaines conjonctions (لأن, إذا, etc.) ocurrent au début des segments pour exprimer une justification, une explication, une hypothèse ou une conclusion.

Les verbes sont aussi employés pour défendre une idée et pour convaincre l'audience de son point de vue. Par exemple, les verbes "اعتقد/croire", "ظن/penser", "توقع/s'attendre", "رأى/voir" expriment des actes d'opinion énoncés par le locuteur pour renforcer ses convictions.

Certains POS sont indicatifs de la nature des actes à assigner. En effet, l'étiquette "Part\_Neg" associée aux mots négatifs ("لا/non", "لم/ne..pas", ...) est utilisée pour exprimer une négation ou un rejet d'une proposition. Aussi, l'interjection "نعم/oui" occure souvent au début des actes d'acceptation d'un point de vue.

Le tableau 3.3 illustre les POS de certains mots clés favorisant la reconnaissance automatique des actes de dialogue appropriés.

Mots clés	POS	Acte correspondant
<i>Mots de question</i>		
كيف/comment	adv_interrog	Demande explication
لماذا/pourquoi	adv_interrog	Demande justification
لم/pourquoi	adv_interrog	Demande justification
هل/est ce que	part_interrog	Question
ماذا/ que	Pron_interrog	Question
ماهو/ quel est	Pron_interrog	Question
ما/qu'est-ce que	Pron_interrog	Question
<i>Conjonctions</i>		
لأن/parce que	Conj_sub	Justification
لكن/mais	conj	Rejet
إذا/si	conj	Hypothèse

Tableau 3.3: Exemples de POS des mots clés

### 6.1.2.3. Descripteurs structurels

Les caractéristiques d'ordre structurel tels que le nom et le rôle du locuteur sont exploitées pour prédire les actes de dialogue correspondants aux segments de dialogue.

Le nom du locuteur peut aider à identifier l'acte correspondant car il peut refléter l'appartenance politique ainsi que les attitudes de l'invité envers certaines thématiques.

Le rôle du locuteur peut intervenir aussi dans l'identification des actes appropriés. En effet, le locuteur peut être présent au débat en tant qu'animateur qui préside la discussion ou en tant qu'invité participant au déroulement de la discussion.

L'animateur est généralement la personne qui gère l'enchaînement des tours de parole entre participants à travers les actes de gestion des obligations sociales (*som*), de demandes (*req*) et de gestion des tours de parole (*tm*).

– *Gestion des obligations sociales (SOM):*

L'animateur débute la discussion par les salutations (A1). Ensuite, il introduit la thématique à discuter (A2) avant de présenter les invités de l'émission (A3). Aussi, il remercie les participants lors de leurs interventions (A4) et clôture la discussion à la fin (A5). Les actes A1, A2, A3, A4 et A5 sont illustrés dans le tableau 3.4.

<b>A1:</b> أهلا بكم في حديث الثورة / Bienvenue au discours de la révolution
<b>A2:</b> يقف التونسيون اليوم على بعد أيام قليلة من قطف أول ثمار ثورتهم التي فجرت الربيع العربي: <i>les Tunisiens se tiennent aujourd'hui sur quelques jours de cueillir les premiers fruits de leur révolution qui a déclenché le printemps arabe.</i>
<b>A3:</b> ولمناقشة هذه القضية ينضم إلينا من تونس السيد العجمي الوريمي: <i>Pour discuter de cette question, nous sommes rejoints par M. Ajmi Ouerimi de la Tunisie.</i>
<b>A4:</b> نشكرك على كل حال أحمد شاكر الناشط الشبابي: <i>Nous vous remercions quand même, Ahmed Shaker, jeune militant.</i>
<b>A5:</b> بهذا تنتهي حلقة اليوم من برنامج حديث الثورة: <i>Ceci conclut l'épisode d'aujourd'hui du programme discours de la révolution</i>

Tableau 3.4: Exemples d'actes de gestion des obligations sociales

– *Demande:*

Le présentateur de l'émission prend l'initiative dans l'animation de la discussion par poser des questions aux invités, demander des explications, des clarifications, des justifications et des opinions sur les thèmes traités tout au long de la discussion. Des exemples de la catégorie demande sont présentés dans le tableau 3.5.

<b>A1:</b> نبدا معك السيد الهاشمي الطرودي، هل تعتقد بأن هذه المدة المخصصة الآن للحملة الانتخابية في تونس كافية؟ <i>Monsieur Hashemi Taroudi, pensez-vous que ce moment est maintenant suffisant pour la campagne électorale en Tunisie?</i>
<b>A2:</b> أستاذ الهاشمي، وكأنك تقول بأن هذه الثورة التي وحدت الشعب التونسي جاءت الأحزاب السياسية بتياراتها المختلفة وأيديولوجيتها لتفرقه؟ <i>Professeur Hachemi, comme si vous disiez que cette révolution qui a uni le peuple tunisien, les partis politiques aux courants et idéologies différents sont venus pour le diviser?</i>
<b>A3:</b> سيد أحمد، ألا تخشى أن تنتكس تونس ثانية إلى حيث كانت؟ <i>Monsieur Ahmed, ne craignez-vous pas que la Tunisie revienne à son niveau actuel?</i>
<b>A4:</b> لماذا المال السياسي ما زال يشكل خطرا على توجهات الناخبين؟ <i>Pourquoi l'argent politique est-il toujours une menace pour les électeurs?</i>

Tableau 3.5: Exemples d'actes de demande

– *Gestion des tours de parole (TM):*

L'animateur est responsable aussi du bon déroulement de l'émission à travers savoir gérer les tensions et les conflits entre les différents participants en cas de désaccord sur une problématique donnée (A1). Donc l'animateur doit apaiser de telles situations pour assurer l'enchaînement du débat dans de bonnes conditions. Aussi, c'est lui qui énonce les pauses et les publicités (A2). Le tableau 3.6 illustre les exemples A1 et A2 d'actes de gestion de tours.

<b>A1:</b> من فضلك سيدي بدون تجريح، نحترم الرأي الآخر.
<i>S'il vous plaît monsieur sans diffamation, nous respectons l'opinion de l'autre.</i>
<b>A2:</b> عند هذه النقطة نتوقف في فاصل قصير ثم نعود لنطرح السؤال التالي.
<i>À ce stade, nous nous arrêtons à un intervalle court et revenons à la question suivante.</i>

Tableau 3.6: Exemples d'actes de gestion de tours de parole

**6.1.2.4. Descripteurs contextuels**

Le contexte d'énonciation est important pour déterminer l'acte actuel à travers l'historique du dialogue. Ainsi l'acte précédent peut donner une idée sur l'acte à prédire.

En effet, la structure de la discussion incorpore souvent des actes de dialogue couplés sous forme de paires d'adjacence à savoir:

- *Question / Réponse:* une réponse est généralement associée à une question qui la précède.
- *Demande d'opinion / Opinion:* une demande d'opinion est généralement suivie par une opinion du participant illustrant son attitude envers une thèse proposée par l'animateur.
- *Thèse / Acceptation:* une thèse est une proposition énoncée par l'animateur qui peut être suivie soit d'une acceptation ou d'un rejet par les intervenants dans le débat.

La liste des paires d'actes de dialogue est détaillée dans le tableau 3.7.

Paire d'adjacence	Exemple
Question/Réponse	<b>A1:</b> سيدة سهام بن سدرين في تونس هل تسمعيني من فضلك؟ <b>A2:</b> أسمعك يا غسان.
Demande explication/Explication	<b>A1:</b> <b>بمعنى؟</b> <b>A2:</b> <b>بمعنى</b> أنه بالتنسيق أو بإيحاء من أميركا هذا اللي حصل.
Demande justification/Justification	<b>A1:</b> <b>لماذا</b> يولد كلام الفريق السيسي هذه المحاذير عندما يتكلم عن تحصيلين المؤسسة العسكرية؟ <b>A2:</b> <b>لأنه</b> يجعل من نفسه ومن القوات المسلحة، القوات المسلحة مؤسسة محترمة من مؤسسات الدولة وليست مؤسسة فوق الدولة،
Demande opinion/Opinion	<b>A1:</b> <b>رأيك</b> سيد سفر؟ <b>A2:</b> <b>أعتقد</b> أميركا لا تحتاج لتركيا حتى تضغط على ننتياهو
Demande confirmation/Confirmation	<b>A1:</b> يعني هو يراجع حساباته؟ <b>A2:</b> <b>طبعاً.</b>

Opinion/Acceptation	A1: لكن لحد الآن المعالجة تبدو أمينة بحتة: A2: صحيح إلى حد كبير
Opinion/Rejet	A1: على كل حال هذه قضية جزئية سيتم التفاوض بشأنها: A2: لا تبدو جزئية:

Tableau 3.7: Exemples de paires d'adjacence

Il est aussi important de connaître le locuteur de l'acte précédent afin de faciliter l'identification de l'acte en cours. Si le locuteur de l'acte précédent est lui-même (A3) donc il peut être en train de poursuivre l'enchaînement de ses idées. Si le locuteur précédent est différent, il s'agit soit de l'animateur (A4) qui pose une question, demande une opinion, une explication ou une justification soit d'un autre participant (A5) qui attend une confirmation ou un rejet de son point de vue. Le tableau 3.8 détaille les exemples de paires P1, P2 et P3 avec même locuteur ou locuteurs différents.

Paire	Locuteur	Exemple
P1: Opinion/ Justification	Même locuteur participant	منى كريم: أعتقد شخصياً أن شروط النجاح متوفرة منى كريم: لأن هذا النجاح يقف على تمثيل أكبر من التيارات السياسية ومن شرائح المجتمع.
P2: Demande explication/Explication	Locuteurs différents Animateur/Participant	ليلي الشايب: هل يمكن أن تفسر ما الذي اشتكى منه سواء الناخبون أو المرشحون؟ شفيق صرصار: هناك إشكال من عدم حياد بعض الهيئات على سبيل المثال.
P3: Opinion/ acceptance	Locuteurs différents Participant1 /Participant2	شفيق صرصار: هذه للافتراضية وليست للواقع فتحجراي: الافتراضية بطبيعة الحال في هذا الإطار

Tableau 3.8: Exemples de paires d'actes avec même locuteur ou locuteurs différents

### 6.1.3. Apprentissage

L'apprentissage automatique consiste à étiqueter correctement des exemples déjà annotés préalablement par un expert. L'objectif est de s'approcher des étiquettes réelles et de fournir un modèle d'apprentissage capable de prédire les étiquettes des nouveaux exemples à annoter. Ainsi, nous avons choisi la technique d'apprentissage supervisé basée sur les modèles CRF afin d'annoter automatiquement les actes de dialogue.

Les CRF se sont montrés efficaces sur de nombreuses tâches d'annotation; ils permettent de générer un ensemble de règles d'apprentissage à partir du corpus d'étude et de la liste des descripteurs favorisant la construction d'un modèle d'apprentissage capable d'annoter de nouvelles étiquettes.

CRF est un modèle graphique discriminant qui exprime des dépendances entre les mots des actes de dialogue à étiqueter. Ce modèle a pour objectif de définir une distribution de probabilités sur les annotations  $a_i$  étant donné un mot  $m_i$  afin de chercher l'annotation la plus probable correspondante.

Ce modèle consiste ainsi à maximiser la probabilité qu'un mot  $m_i$  soit annoté avec l'acte  $a_i$  sachant un ensemble de descripteurs (*Features*)  $f_k$ , ce qui revient à déterminer le poids  $\lambda_k$  associé à chaque descripteur  $f_k$ .

Ainsi, le modèle CRF détermine l'acte  $a_i$  qui maximise la probabilité associé  $P(a_i | m_i)$  pour tous les descripteurs lexicaux, morpho-syntaxiques, structurels et contextuels.

Les classes des actes de dialogue à annoter sont issues du schéma d'annotation SADA19. Ainsi pour chaque acte formé d'un ensemble de mots, on définit deux étiquettes présentant le mot de début (*B-Act*) ainsi que les autres mots inclus dans l'acte (*I-Act*).

Donc le problème revient à classer les mots d'un énoncé dans 38 catégories d'actes de dialogue.

La liste des classes d'actes de dialogue est détaillée dans le tableau 3.9.

N°	Acte de dialogue	Etiquette	Classe	
1	Gestion des obligations sociales	Som	B-Som	I-Som
2	Gestion des tours de parole	Tm	B-Tm	I-Tm
3	Question	Q	B-Q	I-Q
4	Demande de confirmation	Conf_req	B-Conf_req	I-Conf_req
5	Demande d'explication	Expl_Req	B-Expl_Req	I-Expl_Req
6	Demande de justification	Justif_Req	B-Justif_Req	I-Justif_Req
7	Demande d'opinion	Op_Req	B-Op_Req	I-Op_Req
8	Thèse	Th	B-Th	I-Th
9	Opinion	Op	B-Op	I-Op
10	Acceptation	Acc	B-Acc	I-Acc
11	Rejet	Rej	B-Rej	I-Rej
12	Argument	Arg	B-Arg	I-Arg
13	Justification	Justif	B-Justif	I-Justif
14	Explication	Expl	B-Expl	I-Expl
15	Confirmation	Conf	B-Conf	I-Conf
16	Conclusion	Conc	B-Conc	I-Conc
17	Hypothèse	Hyp	B-Hyp	I-Hyp
18	Réponse	Ans	B-Ans	I-Ans
19	Autres	oth	B-oth	I-oth

**Tableau 3.9:** Liste des classes d'actes de dialogue du schéma SADA19

L'entrée étant les énoncés de la discussion déjà segmentés dans l'étape précédente; les mots de chaque énoncé sont déjà annotés par les étiquettes **B** (Begin) et **I** (Inside). L'objectif est

d'assigner à chaque énoncé l'acte de dialogue approprié c'est à dire d'annoter chaque mot  $m_i$  par l'étiquette  $B-Act_i$  ou  $I-Act_i$ .

## 6.2. Phase de test

Cette phase consiste à annoter des nouvelles discussions segmentées au préalable en énoncés. Ceci est assuré en appliquant le modèle d'apprentissage entraîné sur les actes du corpus d'étude utilisé lors de la phase d'apprentissage. La figure 3.16 illustre le détail de cette phase étant donné que l'entrée est une discussion segmentée en énoncés qui sera annotée en actes de dialogue à partir du modèle de classification.

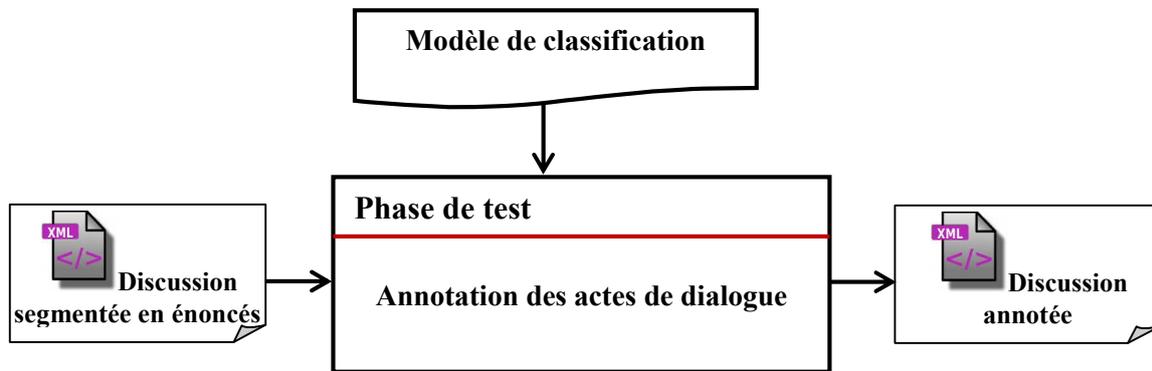


Figure 3.16: Phase de test

La sortie étant une discussion segmentée et annotée en actes de dialogue selon un schéma d'annotation défini en entrée. Un extrait d'une discussion annotée est détaillé dans la figure 3.17.

```

<Turn ID="79" Speaker="غسان بن جدو">
  <Utterance ID="50" DA="استفهام" Class="الإنتشاء الطلبي" Tag="q">؟ أين </Utterance>
</Turn>
<Turn ID="80" Speaker="راشد الغنوشي">
  <Utterance ID="51" DA="إجابة" Class="إجابة" Tag="ans"> موجودة في
  الدستور </Utterance></Turn>
<Turn ID="81" Speaker="غسان بن جدو">
  <Utterance ID="52" DA="قبول" Class="إدارة الحوار" Tag="tm">طيب </Utterance>
  <Utterance ID="53" DA="إلتماس" Class="إدارة الإلتزامات الاجتماعية" Tag="som"> عفوا شيخ راشد
  </Utterance>
  <Utterance ID="54" DA="طلب تفسير" Class="الإنتشاء الطلبي" Tag="expl_req"> يعني هل أفهم من ذلك
  الآن عندما تقول بهذه الطريقة إنك الآن تدعوا إلى تعديل الدستور أو إلغاء الدستور أو ماذا ؟
  </Utterance>
</Turn>
  
```

Figure 3.17: Extrait d'une discussion annotée

## 7. Conclusion

Dans ce chapitre, nous avons détaillé les différentes étapes de la méthode proposée pour la reconnaissance automatique des actes de dialogue dans les discussions arabes.

Notre méthode se base principalement sur une technique d'apprentissage supervisé qui utilise des critères linguistiques et contextuels. Nous avons opté pour le modèle d'apprentissage CRF qui a l'avantage d'être rapide à la fois au niveau de l'étape d'entraînement et de test et d'être performant dans la classification des données séquentielles qui s'applique à la séquence de mots sous forme de texte ou de dialogue. Ce modèle a été appliqué pour les deux tâches de segmentation et d'annotation des actes de dialogue.

Nous détaillons dans le chapitre 4 les résultats expérimentaux qui confirment le choix du modèle CRF en comparant sa performance à d'autres classificateurs. Nous présentons aussi le système de reconnaissance d'actes de dialogue que nous avons développé dans le cadre du présent travail.

---

## **CHAPITRE 4**

### **Implémentation et évaluation du système RECAARD**

---

## 1. Introduction

Rappelons que la méthode de reconnaissance automatique des actes de dialogue que nous avons proposé au niveau du chapitre 3 se base principalement sur une approche numérique et repose plus précisément sur un apprentissage supervisé. Ainsi, nous avons profité des avantages des modèles CRF fréquemment utilisés dans les tâches d'étiquetage des données séquentielles pour implémenter un système de reconnaissance automatique des actes de dialogue.

Le présent chapitre décrit, dans un premier temps, l'architecture générale de notre système nommé RECAARD (**RE**Cognition of **Act** in **AR**abic **D**ebates) et détaille ses principaux modules.

Nous décrivons, dans un deuxième temps, les mesures d'évaluation adoptées pour comparer les différentes techniques d'apprentissage utilisées. Puis, nous discutons les résultats d'évaluation obtenus par notre classificateur en comparaison avec les réseaux bayésiens naïfs (Naive Bayes), les machines à vecteurs de support (SVM) et les arbres de décision (J48). Les différentes expérimentations ont été menées sur les trois versions différentes du corpus de débats CARD (**C**orpus of **AR**abic **D**ebates).

Enfin, nous clôturons ce chapitre par une discussion qui illustre des cas d'erreurs dans le processus d'identification des actes de dialogue.

## 2. Présentation générale du système RECAARD

La méthode de reconnaissance automatique des actes de dialogue proposée au niveau du chapitre 3 a été développée dans le cadre du système RECAARD (**RE**Cognition of **Act** in **AR**abic **D**ebates). Ce système comporte trois principaux modules à savoir le prétraitement des discussions, la segmentation en tours et en énoncés ainsi que l'annotation des actes de dialogue.

Le module de prétraitement consiste à filtrer le débat des mots ambigus et des disfluences afin de fournir un texte en MSA susceptible d'être traité par les outils d'analyse morpho-syntaxique, syntaxique et sémantique généralement développés pour la langue arabe standard.

La discussion prétraitée sera par la suite segmentée en tours délimités par le changement de locuteur. Chaque tour sera par la suite décomposé en énoncés formant l'unité fonctionnelle des segments de dialogue.

Le module d'annotation prend en entrée une discussion segmentée qui sera annotée en actes de dialogue. Ainsi, à chaque énoncé est assigné un acte de dialogue correspondant.

La figure 4.1 illustre l'enchaînement de ces trois modules dans le processus de reconnaissance des actes de dialogue. L'entrée étant une discussion brute sous forme textuelle transcrite à partir d'une émission télévisée.

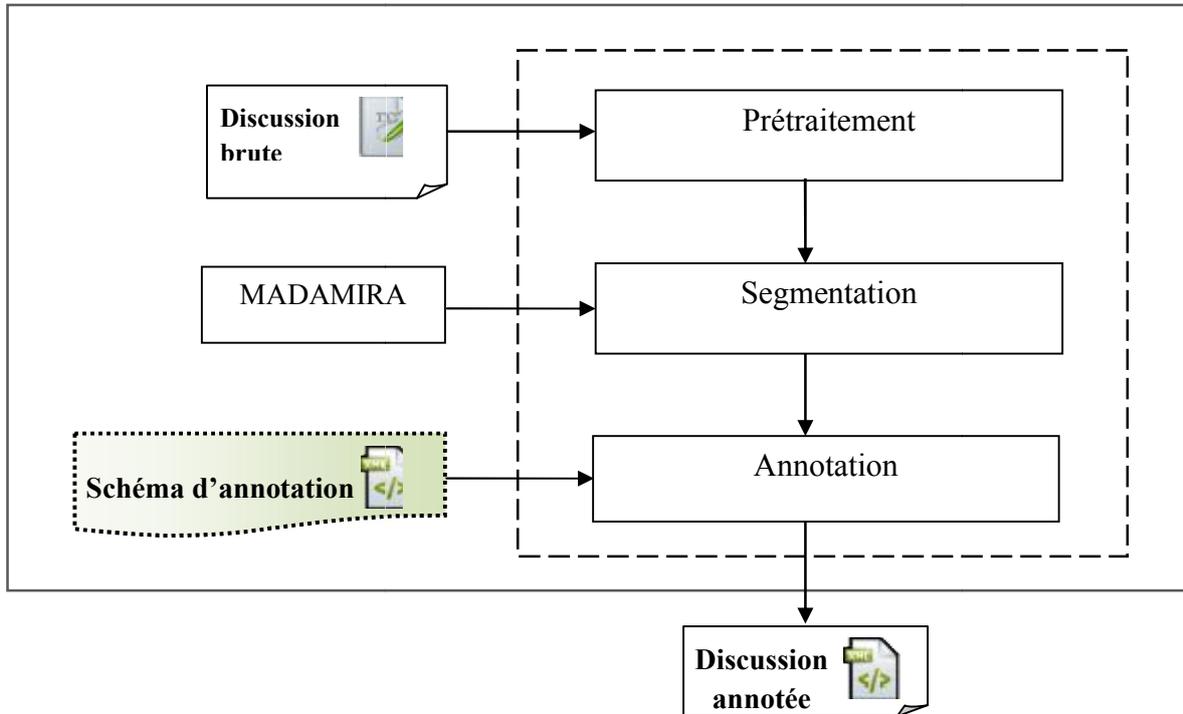


Figure 4.1: Modules du système RECAARD

### 3. Implémentation du système RECAARD

L'implémentation du système RECAARD consiste à développer les modules de prétraitement, de segmentation et d'annotation des actes de dialogue. Nous détaillons dans ce qui suit le processus de réalisation de chaque module.

#### 3.1. Module de prétraitement

Ce module implémente les deux sous étapes de traitement de l'entête et des mots ambigus.

##### 3.1.1. Traitement de l'entête

Étant donné qu'une discussion est formée de deux principales parties à savoir l'entête et le contenu, nous procédons en une première étape à extraire la partie entête du débat "head" qui inclut les éléments suivants:

- Le nom du programme télévisé (exemple حديث الثورة),
- Le sujet à discuter,
- La date de diffusion de l'émission,
- Le nom du présentateur de l'émission,

- La liste des invités au débat,
- Les sous thèmes à traiter tout au long du débat.

Ces informations sont enregistrées dans une balise de début présentée dans la première partie structurant la discussion. Un exemple d'une entête traitée est détaillé dans la figure 4.2.

```
<?xml version="1.0" encoding="UTF-8"?>
<DISCUSSION>
<head>
<program_name>حديث الثورة</program_name>
<episode_name>نزع كيميائي سوريا وخارطة طريق تونس</episode_name>
<animator>محمد كريشان</animator>
<participants>
<participant id="0">عمر إدلبي</participant>
<participant id="1">ريتشارد وايتز</participant>
<participant id="2">أبو هادي الجوراني</participant>
<participant id="3">يوعلي مباركي</participant>
<participant id="4">رفيق عبد السلام</participant>
<participant id="5">محمد جمور</participant>
</participants>
<date>2013/09/12</date>
<subtopics>
<subtopic id="0">اختصار الأزمة السورية بالكيميائي</subtopic>
<subtopic id="1">مجلس الأمن وتعويم القرار</subtopic>
<subtopic id="2">المعارضة المسلحة وتسهيل عمل المراقبين</subtopic>
<subtopic id="3">تونس وبوادر انفراج الأزمة</subtopic>
<subtopic id="4">موافقة النهضة على مبادرة الرباعي</subtopic>
<subtopic id="5">أزمة ثقة بين الفرقاء السياسيين التونسيين</subtopic>
</subtopics>
</head>
```

Figure 4.2: Exemple d'entête d'une discussion

### 3.1.2. Traduction des mots ambigus

Cette sous étape consiste à traduire les expressions dialectales et les mots étrangers empruntés d'autres langues en Arabe Standard Moderne (MSA) afin de faciliter leur analyse dans les étapes ultérieures.

Le processus de traduction se fait d'une manière automatique en se basant sur une liste de mots étrangers et leurs correspondants en arabe. Cette liste inclut 36 mots étrangers et 1248 mots dialectaux. Un extrait de la liste des mots ambigus est présenté dans la figure 4.3.

```
<?xml version="1.0" encoding="UTF-8"?>
<TRANSLATED_W>
<W ID="1" TRANS="حسنا" >ok</W>
<W ID="2" TRANS="نعم" >yes</W>
<W ID="3" TRANS="شكرا" >merci</W>
<W ID="4" TRANS="حسنا" >d'accord</W>
<W ID="5" TRANS="نعم" >oui</W>
<W ID="6" TRANS="لا" >non</W>
<W ID="7" TRANS="أبدا" >pas du tout</W>
<W ID="8" TRANS="أسف" >sorry</W>
<W ID="9" TRANS="أهلا" >hi</W>
<W ID="10" TRANS="أهلا" >hello</W>
<W ID="11" TRANS="سامحتني" >excuse me</W>
<W ID="12" TRANS="كثير" >wide</W>
<W ID="13" TRANS="شكرا" >thanks</W>
<W ID="14" TRANS="كثير" >برشا</W>
<W ID="15" TRANS="كثير" >ياسر</W>
<W ID="16" TRANS="شكرا" >يعيشه</W>
<W ID="16" TRANS="يجب" >لازم</W>
<W ID="16" TRANS="جميل" >بها هي</W>
<W ID="17" TRANS="أين" >وين</W>
<W ID="18" TRANS="لماذا" >علاش</W>
<W ID="19" TRANS="كيف" >كيفاش</W>
<W ID="20" TRANS="لكن" >بس</W>
<W ID="21" TRANS="ليس" >موش</W>
<W ID="22" TRANS="ليس" >مو</W>
<W ID="23" TRANS="طيب" >كويس</W>
</TRANSLATED_W>
```

Figure 4.3: Extrait des mots ambigus traduits en MSA

### 3.2. Module de segmentation

La discussion prétraitée est segmentée en une première étape en tours de parole. Chaque tour est introduit par un nouveau locuteur.

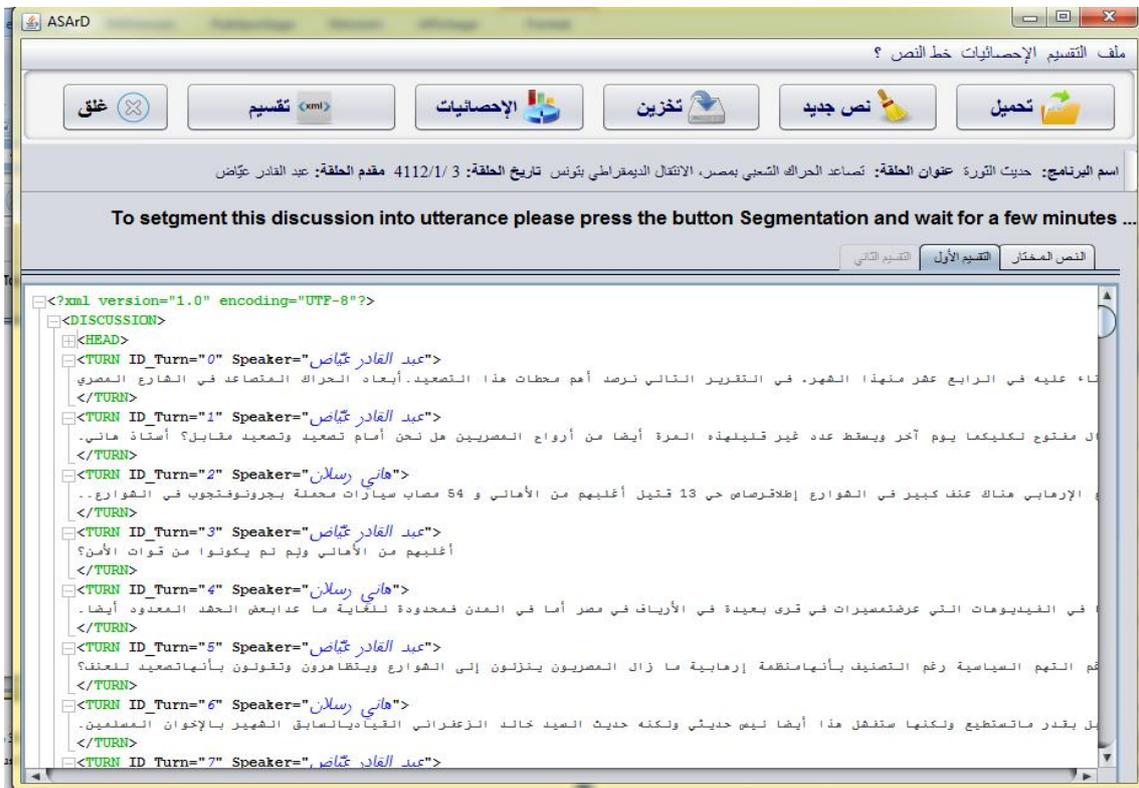


Figure 4.4: Extrait d'une discussion segmentée en tours de parole

Ainsi la segmentation en tours consiste à identifier les changements de locuteur qui sont généralement repérés par la présence du nom du participant suivi de l'indice lexical ":" au début de la ligne. La figure 4.4 illustre un extrait d'une discussion segmentée en tours de parole à travers la balise "Turn" qui inclut la position du tour dans la discussion, le contenu ainsi que l'intervenant qui l'a énoncé.

Au fur et à mesure, la discussion est segmentée en mots; pour chaque mot est associé le lemme et la catégorie grammaticale (POS) correspondants. En effet, la lemmatisation et l'analyse morpho-syntaxique des mots sont assurées par l'analyseur morphologique MADAMIRA (Pasha et al., 2014) qui est une combinaison de deux systèmes implémentés auparavant à savoir MADA (Habash et al., 2013) et AMIRA (Diab et al., 2007).

Cet analyseur est une version améliorée de ces deux systèmes qui permet d'effectuer les tâches de lemmatisation, d'étiquetage morpho-syntaxique et d'analyse morphologique des textes arabes. Implicitement, Madamira permet la segmentation de la discussion en mots (tokenization). Par exemple le mot "السلام" a pour lemme le mot "سلام" et sa catégorie grammaticale (POS) est "noun" puisqu'il s'agit d'un nom.

La figure 4.5 présente un extrait d'une analyse morphologique d'une discussion via l'analyseur morphologique Madamira.

```
<?xml version="1.0" encoding="UTF-8"?>
<DISCUSSION>
<W ID="1" LEMMA="سلام" POS="noun" >السلام</W>
<W ID="2" LEMMA="على" POS="prep" >عليكم</W>
<W ID="3" LEMMA="رحمة" POS="noun" >ورحمة</W>
<W ID="4" LEMMA="الله" POS="noun_prop" >الله</W>
<W ID="5" LEMMA="أهل" POS="adj" >وأهل</W>
<W ID="6" LEMMA="ب" POS="prep" >بكم</W>
<W ID="7" LEMMA="في" POS="prep" >في</W>
<W ID="8" LEMMA="برنامج" POS="noun" >برنامج</W>
<W ID="9" LEMMA="حديث" POS="noun" >حديث</W>
<W ID="10" LEMMA="ثورة" POS="noun" >الثورة</W>
<W ID="11" LEMMA="قدم" POS="verb" >ونقدم</W>
<W ID="12" LEMMA="في" POS="prep" >في</W>
<W ID="13" LEMMA="جزء" POS="noun" >جزئه</W>
<W ID="14" LEMMA="أول" POS="adj_num" >الأول</W>
<W ID="15" LEMMA="قراءة" POS="noun" >قراءة</W>
<W ID="16" LEMMA="في" POS="prep" >في</W>
<W ID="17" LEMMA="قرار" POS="noun" >قرار</W>
<W ID="18" LEMMA="مجلس" POS="noun" >مجلس</W>
<W ID="19" LEMMA="أمن" POS="noun" >الأمن</W>
<W ID="20" LEMMA="حول" POS="noun" >حول</W>
<W ID="21" LEMMA="نزاع" POS="noun" >نزاع</W>
<W ID="22" LEMMA="سلاح" POS="noun" >السلاح</W>
<W ID="23" LEMMA="كيميائي" POS="adj" >الكيميائي</W>
<W ID="24" LEMMA="سوري" POS="adj" >السوري</W>
```

Figure 4.5: Extrait de l'analyse morphologique effectuée par Madamira

En une deuxième étape, chaque tour de parole est décomposé en énoncés. En effet, le processus de segmentation en énoncés est réalisé en appliquant la technique d'apprentissage supervisé qui implémente le modèle CRF. Ainsi à chaque mot du segment de dialogue est assigné soit la catégorie *B* si le mot est introduit au début soit la catégorie *I* pour un mot situé à l'intérieur de l'énoncé. Cette information sur la position du mot dans l'énoncé est intégrée dans la balise "Bound" qui prend la valeur *B* ou *I*. Ainsi chaque nouvel énoncé est marqué par la balise "Bound"="B". Un extrait du fichier segmenté en mots détaillant les limites de segments est présenté dans la figure 4.6.

```
<?xml version="1.0" encoding="UTF-8" ?>
<DISCUSSION>
<W ID="1" LEMMA="سلام" POS="noun" BOUND="B">السلام</W>
<W ID="2" LEMMA="علي" POS="prep" BOUND="I">عليكم</W>
<W ID="3" LEMMA="رحمة" POS="noun" BOUND="I">ورحمة</W>
<W ID="4" LEMMA="الله" POS="noun_prop" BOUND="I">الله</W>
<W ID="5" LEMMA="أهل" POS="adj" BOUND="B">وأهل</W>
<W ID="6" LEMMA="ب" POS="prep" BOUND="I">بكم</W>
<W ID="7" LEMMA="في" POS="prep" BOUND="I">في</W>
<W ID="8" LEMMA="برنامج" POS="noun" BOUND="I">برنامج</W>
<W ID="9" LEMMA="حديث" POS="noun" BOUND="I">حديث</W>
<W ID="10" LEMMA="ثورة" POS="noun" BOUND="I">الثورة</W>
<W ID="11" LEMMA="قدم" POS="verb" BOUND="B">ونقدم</W>
<W ID="12" LEMMA="في" POS="prep" BOUND="I">في</W>
<W ID="13" LEMMA="جزء" POS="noun" BOUND="I">جزئه</W>
<W ID="14" LEMMA="أول" POS="adj_num" BOUND="I">الأول</W>
<W ID="15" LEMMA="قراءة" POS="noun" BOUND="I">قراءة</W>
<W ID="16" LEMMA="في" POS="prep" BOUND="I">في</W>
<W ID="17" LEMMA="قرار" POS="noun" BOUND="I">قرار</W>
<W ID="18" LEMMA="مجلس" POS="noun" BOUND="I">مجلس</W>
<W ID="19" LEMMA="أمن" POS="noun" BOUND="I">الأمن</W>
<W ID="20" LEMMA="حول" POS="noun" BOUND="I">حول</W>
<W ID="21" LEMMA="نزاع" POS="noun" BOUND="I">نزاع</W>
<W ID="22" LEMMA="سلاح" POS="noun" BOUND="I">السلاح</W>
<W ID="23" LEMMA="كيمياي" POS="adj" BOUND="I">الكيمياي</W>
<W ID="24" LEMMA="سوري" POS="adj" BOUND="I">السوري</W>
```

Figure 4.6: Extrait d'un fichier segmenté en mots et en énoncés

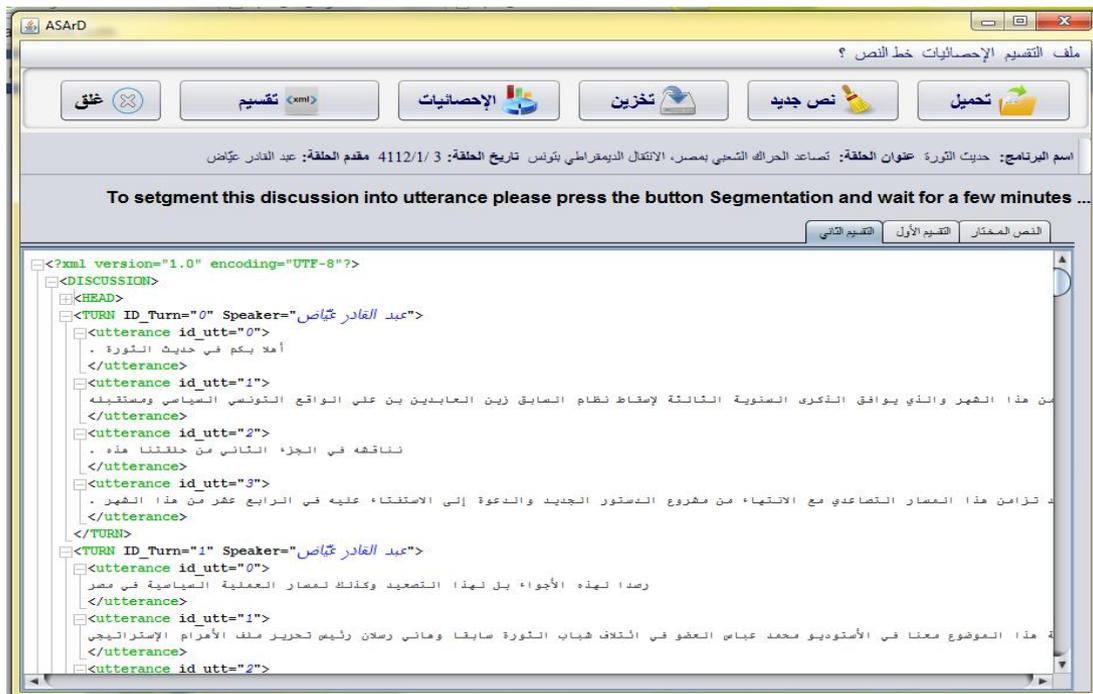


Figure 4.7: Extrait d'une discussion segmentée en énoncés

Ainsi le fichier segmenté en tours illustré dans la figure 4.4 est raffiné en intégrant les balises "utterance" dans chaque balise "turn". Ces nouvelles balises représentent les segments de dialogue inclus dans chaque tour de parole. Le fichier généré lors de l'étape de segmentation est détaillé dans la figure 4.7.

### 3.3. Module d'annotation

Le module d'annotation consiste à assigner un acte de dialogue à chaque énoncé segmenté dans l'étape précédente. Ce module a suscité une première étape d'annotation semi-automatique afin de préparer un corpus annoté. Ce corpus sera utilisé dans une deuxième étape d'annotation automatique par apprentissage.

#### 3.3.1. Annotation semi-automatique

Afin de faciliter la construction d'un large corpus de débats annotés, nous avons développé un outil d'aide à l'annotation nommé "ActAAr" (**Act**Annotation in **Arabic**) (BenDbabis et al., 2012). Cet outil permet principalement à un expert humain d'annoter une discussion selon un schéma d'annotation donné. L'interface principale du système "ActAAr" est bilingue; l'utilisateur peut visualiser les données soit en arabe soit en anglais selon un choix qu'il effectue à travers l'interface d'accueil présentée dans la figure 4.8.



Figure 4.8: Interface d'accueil d'ActAAr

Pour débiter l'annotation, l'utilisateur doit choisir une discussion sous format texte ou html. La discussion sera par la suite affichée dans l'onglet central « Selected Text » de l'interface présentée dans la figure 4.9.

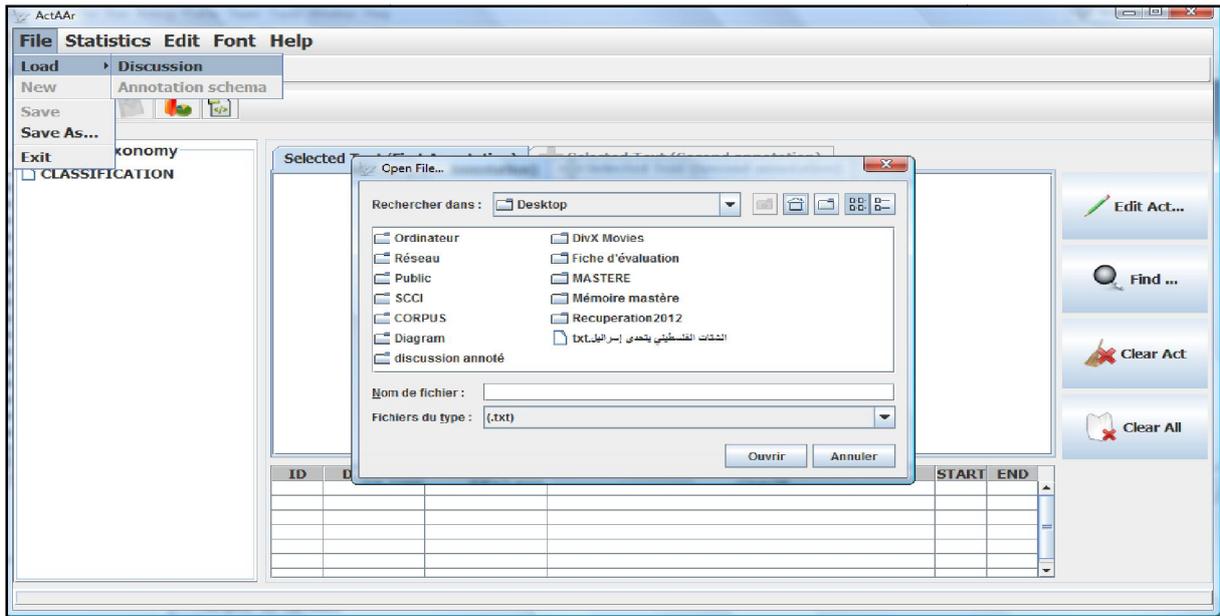


Figure 4.9: Choix de la discussion à annoter

Une fois la discussion est chargée, l'utilisateur doit procéder au choix de la classification d'actes de dialogue. Le schéma d'annotation choisi est affiché sous forme d'une arborescence à gauche de l'interface principale. La figure 4.10 décrit le processus du choix d'un schéma d'annotation d'actes de dialogue.

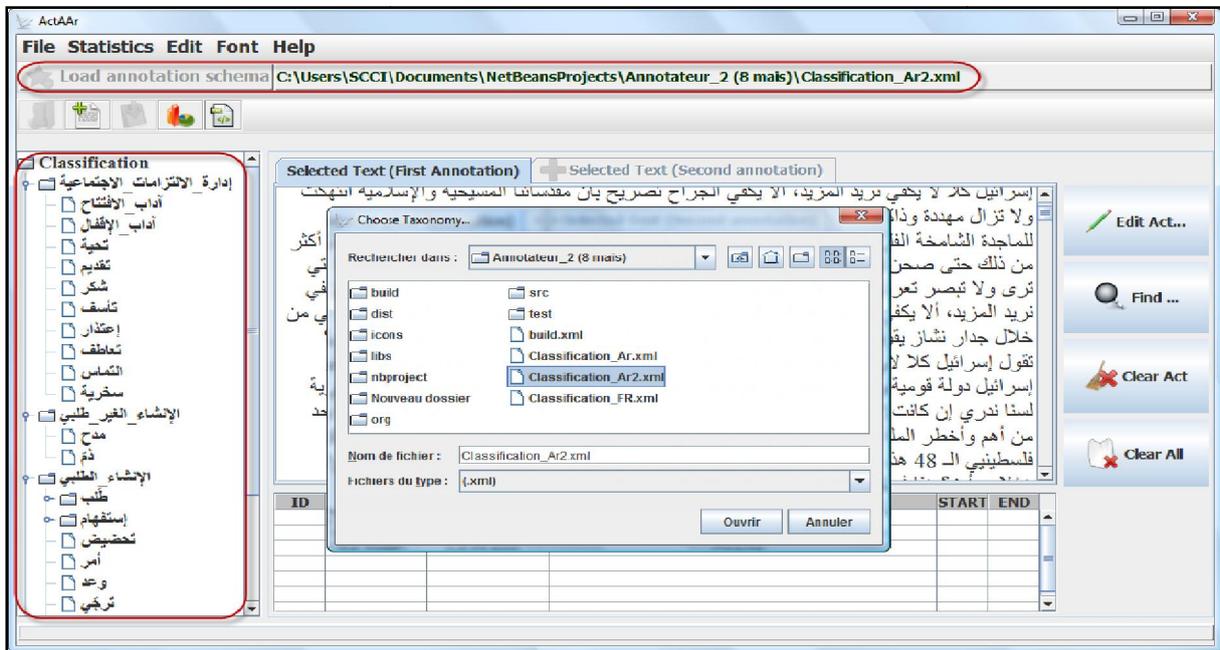


Figure 4.10: Choix du schéma d'annotation

Pour commencer le processus d'annotation, l'utilisateur sélectionne l'énoncé à annoter et choisit l'acte de dialogue approprié à partir de la liste des actes issus du schéma d'annotation

défini auparavant. La figure 4.11 présente le processus d'annotation semi-automatique guidée par le choix de l'utilisateur des étiquettes appropriées à chaque énoncé.

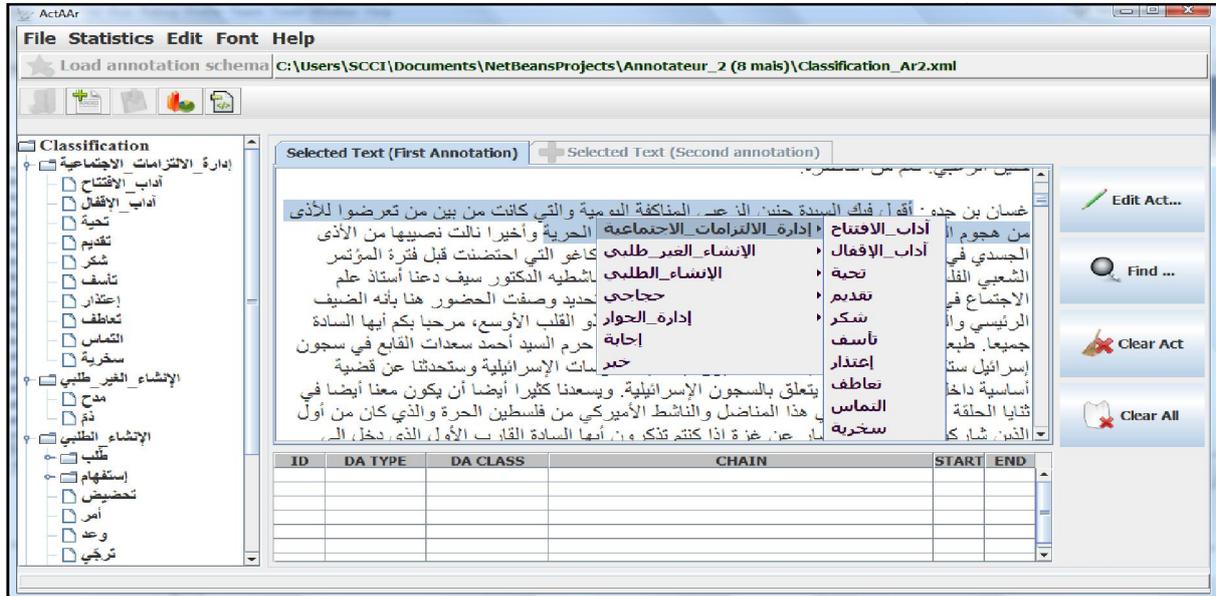


Figure 4.11: Annotation d'un acte de dialogue

Chaque acte annoté sera colorié avec une couleur spécifique pour le distinguer des autres actes inclus dans le schéma d'annotation correspondant. La figure 4.12 illustre une discussion annotée avec un schéma donné.

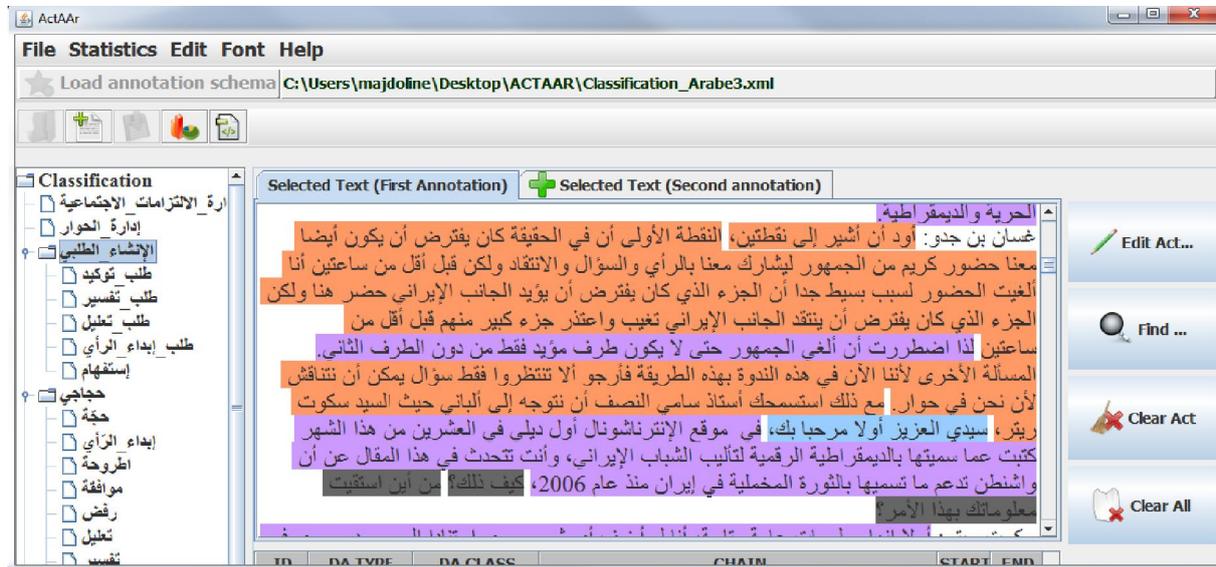


Figure 4.12: Annotation semi-automatique via l'outil ActAAR

L'utilisateur peut modifier l'étiquette d'un acte de dialogue déjà annoté à travers l'interface présentée dans la figure 4.13. Il peut aussi effectuer une recherche des énoncés par acte de dialogue dans l'interface présentée dans la figure 4.14

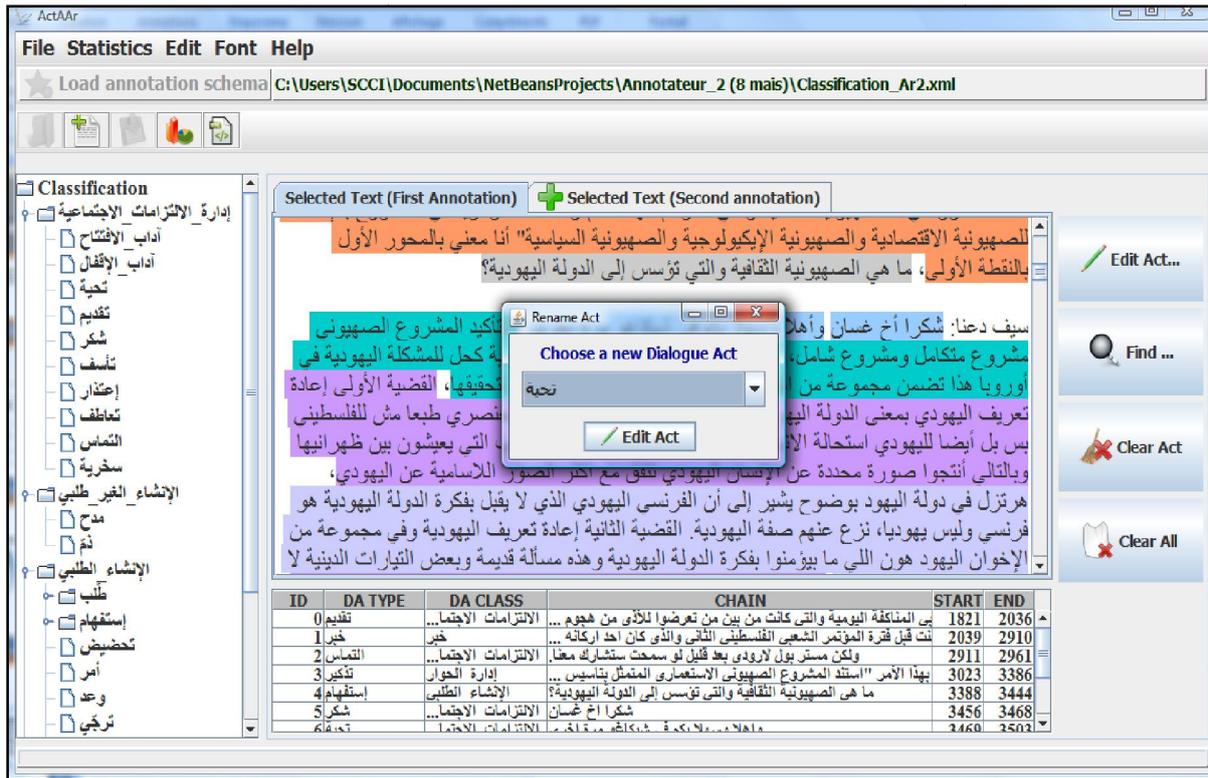


Figure 4.13: Modification d'un acte de dialogue

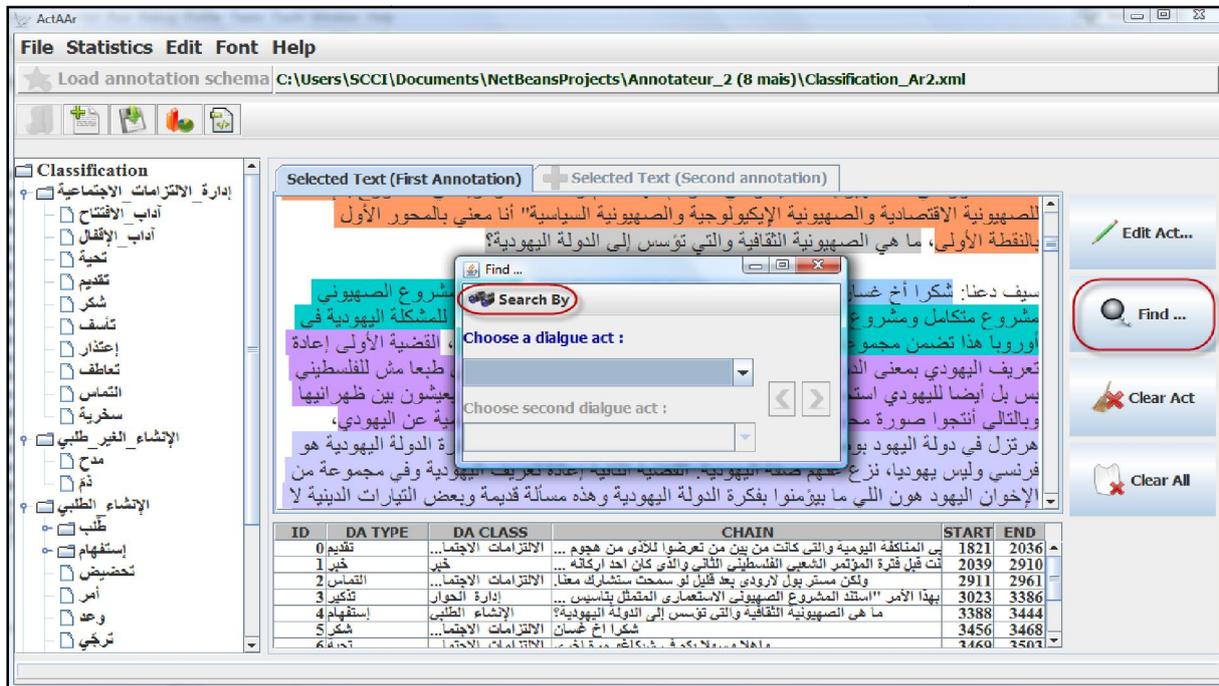


Figure 4.14: Recherche d'un acte de dialogue

L'outil ActAAR permet d'effectuer des statistiques à savoir le calcul de la fréquence de chaque acte dans la discussion annotée. La figure 4.15 illustre les statistiques par catégorie d'actes ainsi que la dispersion des actes dans chaque catégorie.

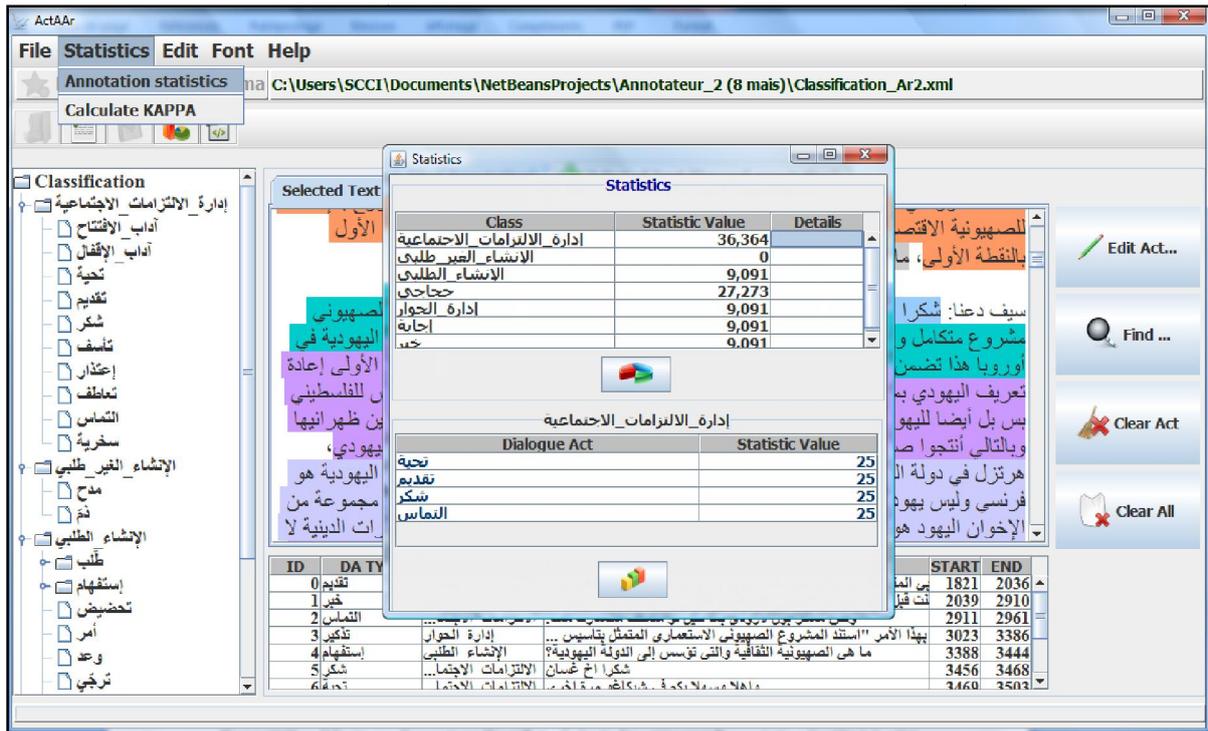


Figure 4.15: Statistiques de l'outil ActAAR

Par ailleurs, l'outil ActAAR permet aussi de mesurer l'accord inter-annotateur à travers le calcul du coefficient KAPPA entre deux experts humains sur l'annotation d'une discussion donnée. L'interface de calcul du ratio kappa est présentée dans la figure 4.16.

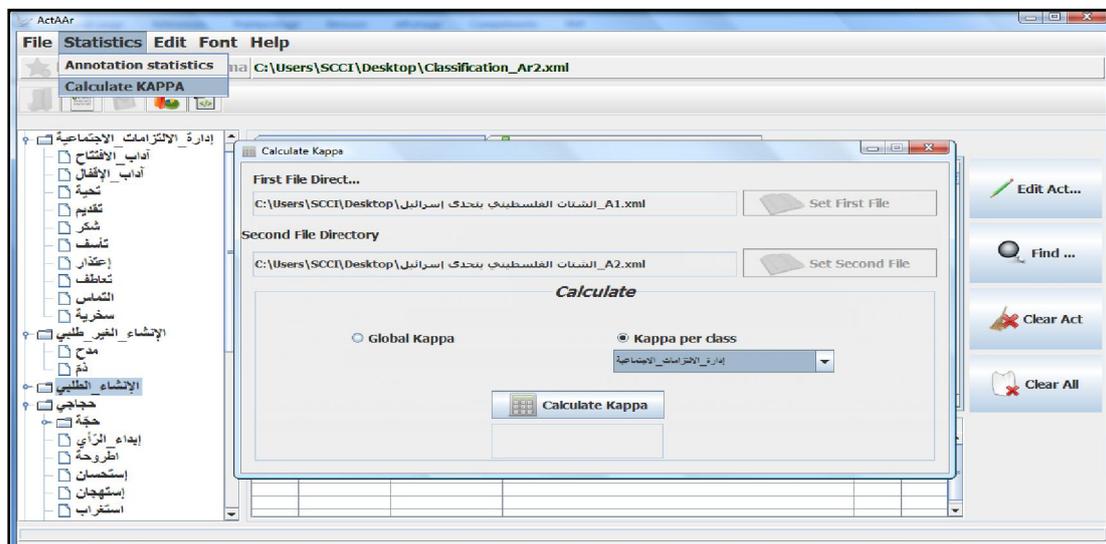


Figure 4.16: Calcul de Kappa via l'outil ActAAR

### 3.3.2. Annotation automatique

Le module d'annotation automatique est implémenté à travers les champs aléatoires conditionnels. En effet, le modèle CRF exploite le corpus de discussions annotées ainsi que la liste des caractéristiques d'apprentissage d'ordre lexical, morpho-syntaxique, structurel et

contextuel afin de générer un modèle d'apprentissage. Ainsi le modèle généré permet d'étiqueter de nouvelles discussions non annotées au préalable. Le fichier résultant du module d'annotation automatique présente la sortie finale du système RECAARD. En effet à chaque énoncé segmenté sera assigné automatiquement un acte de dialogue correspondant. Donc la balise "utterance" du fichier segmenté (sortie de l'étape de segmentation) sera enrichie par les informations suivantes:

- le nom de l'acte de dialogue assigné,
- la catégorie à laquelle appartient l'acte choisi (demande, argumentation, gestion de dialogue, etc.),
- L'étiquette ("tag") de l'acte choisi.

Le fichier résultat du système RECAARD est détaillé dans la figure 4.17.

```
<?xml version="1.0" encoding="UTF-8"?>
<DISCUSSION>
<head>
<turn speaker="محمد كريشان" id_turn="1" length="917">
<utterance name="إدارة_الالتزامات_الاجتماعية" class="إدارة_الالتزامات_الاجتماعية" tag="som" id="0" start="13" end="36" color="1">
السلام عليكم ورحمة الله
</utterance>
<utterance name="إدارة_الالتزامات_الاجتماعية" class="إدارة_الالتزامات_الاجتماعية" tag="som" id="1" start="37" end="69" color="1">
وأملأ بكم في برنامج حديث الثورة
</utterance>
<utterance name="اطروحة" class="حجاجي" tag="th" id="2" start="70" end="195" color="13">
وتقدم في جزئه الأول قراءة في قرار مجلس الأمن حول نزاع السلاح الكيماوي السوري وتداعياته على المعارفة المسلحة في الداخل السوري
</utterance>
<utterance name="إدارة_الالتزامات_الاجتماعية" class="إدارة_الالتزامات_الاجتماعية" tag="som" id="5" start="930" end="1020" color="1">
معنا في هذا الجزء من البرنامج منا في الأستديو عمر إدلبي ممثل تجمع أنصار الثورة السورية
</utterance>
<utterance name="إدارة_الالتزامات_الاجتماعية" class="إدارة_الالتزامات_الاجتماعية" tag="som" id="8" start="1157" end="1178" color="1">
أملأ بغيرقنا جميعاً
</utterance>
<utterance name="طلب_توكيد" class="الإنشاء_الطلبية" tag="conf_req" id="9" start="1179" end="1274" color="6">
سيد عمر إدلبي هل فعلاً قرار مجلس الأمن هو كما قال بان كي مون هو أول بارقة أمل في الملف السوري؟
</utterance>
</turn>
<turn speaker="عمر إدلبي" id_turn="3" length="1080">
<utterance name="توكيد" class="حجاجي" tag="conf" id="10" start="1288" end="1365" color="13">
يعني وفق تصور بان كي مون للحل في سوريا فهو نعم بارقة أمل بالنسبة لبان كي مون
</utterance>
<utterance name="إبداء_الرأي" class="حجاجي" tag="op" id="12" start="1485" end="1666" color="13">
ولكن بالنسبة لنا نعتقد أنه ليس أكثر من محطة جديدة من محطات تعويم هذا النظام من قبل المجتمع الدولي، تقديم مزيد من الفرص له ليتمكن أو ليكون عنده مجال أوسع
للقتضاء على ثورة الشعب السوري
</utterance>
<utterance name="تعلييل" class="حجاجي" tag="justif" id="13" start="1667" end="1759" color="13">
لأن كل ما أتي في هذا القرار لا يمس الشعب السوري بصله أو لا يمت للشعب السوري بصله على الإطلاق
</utterance>
</turn>
```

Figure 4.17:Sortie finale du système RECAARD

## 4. Évaluations

La reconnaissance des actes de dialogue dans les débats télévisés arabes a été réalisée à travers les modèles d'apprentissage CRF. Pour se faire, nous avons préparé un corpus de discussions pour l'apprentissage et la prédiction automatique des limites de segments ainsi que des étiquettes correspondantes à chaque segment. La taille du corpus d'apprentissage inclut 50 discussions contenant 260212 mots et 14062 actes de dialogue. Nous avons appliqué des descripteurs que nous avons jugés pertinents pour les tâches de segmentation et d'annotation des actes de dialogue. Nous détaillons dans ce qui suit les techniques d'évaluation adoptées pour juger la performance des modèles CRF ainsi que résultats expérimentaux obtenus sur les différentes versions du corpus CARD. Nous illustrons et analysons aussi les cas erronés dans la classification des actes de dialogue.

### 4.1. Mesures d'évaluation

Pour évaluer l'efficacité du processus de reconnaissance des actes de dialogue, nous avons choisi de calculer les valeurs des mesures d'évaluation habituellement utilisées pour l'évaluation des résultats d'apprentissage obtenus par un classifieur à savoir la précision (P), le rappel (R) et la F-mesure (F) pour les deux modules de segmentation et d'annotation des actes de dialogue. Le taux d'erreur est mesuré seulement au niveau de l'étape de segmentation. Il mesure la sensibilité du système en matière de classification des données c'est à dire la performance exacte du système. Il est obtenu par la formule suivante :

$$\text{Taux d'erreur} = \frac{\text{nombre d'instances faussement classifiées}}{\text{nombre total d'instances}} \quad (14)$$

La précision est le pourcentage des énoncés correctement étiquetés par rapport au nombre total d'énoncés fournis par le système.

Le rappel présente le rapport entre les énoncés correctement classifiés par le système et le nombre total d'énoncés réellement bien classifiés dans le corpus de référence.

La valeur de f-mesure calcule la moyenne pondérée entre le rappel et la précision selon la formule suivante :

$$F\text{-mesure} = \frac{2 * \text{Précision} * \text{Rappel}}{\text{Précision} + \text{Rappel}} \quad (15)$$

## 4.2. Évaluation du module de segmentation

Afin de valider la tâche de segmentation automatique des débats télévisés en énoncés, nous avons évalué la performance des modèles CRF en comparaison avec d'autres classificateurs en effectuant une série d'expérimentations sur les différentes versions du corpus CARD.

### 4.2.1. Expérimentations

La tâche de segmentation consiste à déterminer pour chaque mot s'il représente un mot au début d'un segment ou s'il se situe à l'intérieur d'un segment.

Pour se faire, nous avons défini deux classes : *B* (*Begin*) pour le premier mot du segment et *I* (*Inside*) pour un mot à l'intérieur du segment. Ainsi la segmentation des discussions en énoncés revient à résoudre un problème de classification binaire (classes B/I).

Afin d'évaluer les performances des modèles CRF, nous avons utilisé plusieurs modèles appliqués au niveau mot, qui diffèrent selon les dépendances locales de chaque mot dans le même énoncé. Ces dépendances varient selon l'intervalle uni-gramme (un mot) ou bi-grammes (deux mots) du mot à prédire.

Suite à une première expérimentation, nous avons limité le nombre de modèles expérimentés à quatre.

Le premier (noté M0) est un modèle qui ne prend en considération aucune dépendance entre les mots d'un même énoncé. Dans ce cas, les modèles CRF jouent le rôle d'un simple étiqueteur.

Le deuxième (noté M1) est un modèle qui exploite une fenêtre de deux mots tenant compte du mot précédent et du mot suivant dans le même énoncé.

Le troisième (noté M2) est un modèle qui utilise une fenêtre mettant en jeu deux mots en aval et deux mots en amont, par rapport au mot courant.

Le quatrième modèle (noté M3) consiste à améliorer le troisième modèle en ajoutant deux dépendances locales. Cette dépendance utilise deux bi-grammes tenant compte du mot courant avec le mot précédent (respectivement le mot suivant).

Ces modèles sont ensuite exploités pour l'apprentissage des paramètres CRF sur la dernière version du corpus CARD (CARD 1.2) qui inclut 260383 mots. Le tableau 4.1 illustre les résultats de l'évaluation de la segmentation utilisant les différents modèles en termes de précision, de rappel, de F-mesure et de taux d'erreur.

	M0	M1	M2	<b>M3</b>
Rappel (%)	85,2	92,1	96,7	<b>98</b>
Précision (%)	84,4	94	96,2	<b>97,8</b>
F-mesure (%)	84,8	93,55	96,45	<b>97,9</b>
Taux d'erreur (%)	12,2	6,6	3,4	<b>2,05</b>

*Tableau 4.1: Évaluation de la segmentation selon différents modèles d'apprentissage*

D'après les résultats obtenus, nous remarquons que le modèle M3 donne les meilleurs résultats ce qui explique l'importance des dépendances entre mots pour définir les limites correctes des segments.

Pour valider l'efficacité du classificateur CRF, nous avons opté à une étude comparative des résultats obtenus par le meilleur modèle (M3) à d'autres classificateurs à savoir :

- SMO : implémente la méthode d'apprentissage vectorielle SVM.
- Naïve Bayes (NB) : implémente le modèle probabiliste des réseaux bayésiens naïfs.
- J48 : implémente le modèle génératif des arbres de décision.

Pour évaluer les différents modèles, nous avons divisé le corpus de discussions collecté en deux parties : 80% pour l'apprentissage et 20% pour le test. La répartition du corpus en nombre de discussions et en mots pour les trois versions du corpus CARD est présentée dans le tableau 4.2.

	CARD 1.0		CARD 1.1		CARD 1.2	
	# discussions	# mots	# discussions	# mots	# discussions	# mots
Apprentissage	6	29660	17	80935	42	208306
Test	2	7415	5	20234	8	52077
Total	8	37075	22	101169	50	260383

*Tableau 4.2: Répartition du corpus de discussions CARD*

Les résultats de l'étude comparative à travers les trois versions du corpus CARD sont illustrés dans le tableau 4.3.

	Rappel			Précision			F-mesure		
	CARD 1.0	CARD 1.1	CARD 1.2	CARD 1.0	CARD 1.1	CARD 1.2	CARD 1.0	CARD 1.1	CARD 1.2
CRF (%)	88,4	97,4	<b>98</b>	87	97,2	<b>97,8</b>	87,7	97,3	<b>97,9</b>
NB (%)	83	94	94,7	82	92	93,2	82,5	93	93,9
SMO (%)	83	94	95	84	93	94,6	83,5	93,5	94,8
J48 (%)	82	94	94,2	81,5	89	90,3	81,7	91,4	92,2

*Tableau 4.3: Résultats d'évaluation de la tâche de segmentation*

Les résultats d'évaluation de la tâche de segmentation montrent que la taille du corpus influence sur les résultats d'apprentissage. En effet, plus le nombre de mots augmente plus les résultats de

segmentation s'améliorent. Ainsi, la dernière version du corpus CARD (CARD 1.2) contenant 260383 mots répartis sur 50 débats télévisés a donné les meilleurs résultats pour tous les classificateurs. Ainsi les résultats d'évaluation augmentent de plus de 10% en passant de la première version CARD 1.0 formé 37075 mots vers le corpus CARD 1.2 ce qui confirme l'importance de la taille du corpus dans le processus d'apprentissage.

L'efficacité des modèles CRF est confirmée en comparaison avec les autres modèles. En effet, CRF a donné une meilleure pondération entre le rappel et la précision pour toutes les versions du corpus CARD. Les meilleurs résultats d'évaluation sont obtenus avec la dernière version CARD 1.2. Ainsi, la valeur de f-mesure (97,9%) dépasse les valeurs obtenues par les autres classifieurs à savoir SMO (f-mesure=94,8%), NB (f-mesure=93,9%) et J48 (f-mesure=92,2%) pour le corpus de débats CARD 1.2.

#### 4.2.2. Analyse des cas d'erreurs

Après une observation des résultats de segmentation obtenus par CRF, nous avons constaté quelques causes d'erreurs dans la détermination des limites des segments d'énoncés à savoir les ambiguïtés causées par les signes de ponctuation, les mots clés ainsi que l'analyse morphologique.

##### - Erreurs dues à la ponctuation

Dans la plupart des cas un signe de ponctuation désigne la fin de l'énoncé c'est-à-dire le mot après un signe de ponctuation est généralement un mot de début d'énoncé qui doit avoir comme classe "B". Dans la langue arabe, on utilise généralement des phrases longues insensibles à la ponctuation donc on peut enchaîner la même idée en utilisant une virgule ou un point-virgule.

#Mot	#POS	#M_Clé	#M_Quest	#Punc	Class_Man	#Class_CRF
ما	pron_rel	1	1	0	B	B
هي	Pron	0	0	0	I	I
حدود	Noun	0	0	0	I	I
مسؤوليتكم	Noun	0	0	0	I	I
،	<b>Punc</b>	0	0	1	I	I
وإلى	<b>Prep</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>I</b>	<b>B</b>
أين	adv_interrog	1	1	0	I	I
ترون	Verb	1	0	0	I	I
الأمر	Noun	0	0	0	I	I
؟	Punc	1	0	1	I	I

Tableau 4.4: Exemple d'une segmentation erronée due à la ponctuation

Ainsi, les marques de ponctuation peuvent se présenter à l'intérieur d'un segment ce qui engendre une fausse segmentation. Nous présentons dans le tableau 4.4 un exemple où la ponctuation cause une segmentation erronée vu la présence d'un signe de ponctuation au milieu du segment.

- Erreurs dues à l'emplacement des mots clés

Les mots clés ne sont pas toujours un critère discriminant pour délimiter les segments de dialogue. En effet, ils ne sont pas toujours présents au début des énoncés, ils peuvent exister comme mots à l'intérieur d'un énoncé. Le tableau 4.5 illustre un exemple contenant un énoncé exprimant une acceptation suivie d'une justification. Ainsi, le mot clé "سببه/sa cause" engendre une fausse segmentation vu qu'il est présent à l'intérieur de l'énoncé:

" نعم هناك تصعيد و هذا التصعيد سببه أن هناك عنفا كبيرا في الشوارع "

#Mot	#POS	#M_Clé	#M_Quest	#Punc	Class_Man	#Class_CRF
نعم	pron_rel	1	0	0	B	B
هناك	Pron	0	0	0	I	I
تصعيد	Noun	0	0	0	I	I
وهذا	<b>Noun</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>B</b>	<b>I</b>
التصعيد	Pron	0	0	0	I	I
سببه	<b>Noun</b>	<b>1</b>	<b>0</b>	<b>0</b>	<b>I</b>	<b>B</b>
أن	conj_sub	0	0	0	I	I
هناك	adv	0	0	0	I	I
عنفا	noun	0	0	0	I	I
كبيراً	adj	1	0	0	I	I
في	prep	0	0	0	I	B
الشوارع	noun	0	0	0	I	I

Tableau 4.5: Exemple d'une fausse segmentation due aux mots clés

### 4.3. Évaluation du module d'annotation

L'évaluation de la tâche d'annotation permet de mesurer la performance des modèles CRF dans l'étiquetage des énoncés segmentés dans l'étape de segmentation. Nous détaillons dans ce qui suit les expérimentations effectuées, les résultats obtenus en utilisant les trois versions du corpus CARD ainsi que l'analyse des cas défaillants.

#### 4.3.1. Expérimentations

Après une première étape de segmentation, les énoncés résultants seront étiquetés par les actes de dialogue correspondants. Rappelons qu'à chaque segment de dialogue correspond un seul acte qui exprime l'intention du participant. Par exemple, l'énoncé *مرحباً بكم /mrHbA bikum/bienvenue* est étiqueté par l'acte *SOM* qui désigne la gestion des obligations sociales.

Pour se faire, nous avons préparé un corpus de débats politiques annotés manuellement par des experts humains. Ce corpus a été construit sur trois étapes vu la complexité de la tâche d'annotation des discussions argumentatives en termes de temps et de coût. Le point de départ était avec 8 discussions (noté CARD 1.0) que nous avons exploitées pour construire une première version du schéma d'annotation SADA40. Ce corpus inclut un nombre limité d'actes de dialogues (2367 actes) qui est insuffisant pour la classification automatique de 40 actes. En un deuxième temps, nous avons élargi le corpus à 22 débats (CARD 1.1) contenant 6050 actes. Cette deuxième version a été annotée avec un schéma d'annotation réduit (noté SADA19) qui se limite à 19 actes exprimant principalement des demandes et des actes argumentatifs (opinion, argument, acceptation, rejet, etc.). Cette nouvelle version a amélioré les résultats d'apprentissage mais reste insuffisante pour prédire la plupart des actes argumentatifs. Ainsi, nous avons enrichi la taille du corpus à 50 discussions qui incluent 14062 actes de dialogue (CARD 1.2) qui ont été annotés avec le schéma SADA19. Cette dernière version a donné des résultats satisfaisants ce qui justifie l'importance de la taille du corpus dans le processus d'apprentissage automatique des données.

Pour l'évaluation du module d'annotation des énoncés segmentés en actes de dialogue, nous avons calculé les mesures de rappel, de précision et de f-mesure.

Pour valider l'efficacité du classificateur choisi, nous avons utilisé la technique de validation croisée avec 10 itérations. Cette technique est généralement adoptée lorsqu'on dispose d'un petit corpus qui ne peut pas être décomposé en deux parties : apprentissage et test. Donc le système décompose automatiquement le corpus collecté et itère la même opération 10 fois.

Ce choix est justifié puisqu'on a commencé l'apprentissage avec une collection de 8 discussions qui inclut seulement 2367 actes de dialogue dont la prédiction est difficile via une taxonomie de 40 classes d'actes. Nous avons gardé la même technique de validation pour les deux autres versions du corpus pour garantir les mêmes conditions d'évaluation.

#### ***4.3.1.1. Résultats de CARD 1.0***

Le modèle CRF a été appliqué sur le corpus CARD 1.0 contenant 2367 actes de dialogue. L'évaluation de ce classificateur est réalisée en termes de rappel, de précision et de f-mesure.

Les résultats d'évaluation de CRF ont été comparés à ceux d'autres classifieurs tels que les arbres de décision (J48), les réseaux bayésiens (NB) et les machines à vaste marge (SMO). Les résultats de l'étude comparative sont détaillés dans le tableau 4.6.

	<i>Correct</i>	<i>Incorrect</i>	<i>Précision</i>	<i>Rappel</i>	<i>F-mesure</i>
<i>J48 (%)</i>	24,98	75,02	24,5	25	24,3
<i>NB (%)</i>	29,54	70,46	35,8	29,5	25,2
<i>SMO (%)</i>	31,7	68,3	33,8	31,7	29,8
<i>CRF (%)</i>	30,07	69,93	35,9	30,1	31,8

*Tableau 4.6: Résultats d'annotation du corpus CARD 1.0*

Les expérimentations réalisées confirment l'efficacité des modèles CRF en comparaison avec les autres classificateurs (F-mesure=32,03%). Cependant les valeurs obtenues montrent des scores faibles de classification des actes de dialogue; c'est-à-dire le modèle d'apprentissage n'a pas pu prédire la plupart des actes. Afin d'étudier les cas d'échec au niveau du processus de classification, nous avons détaillé les résultats d'évaluation par acte de dialogue dans le tableau 4.7.

<i>Acte</i>	<i>Précision</i>	<i>Rappel</i>	<i>F-mesure</i>	<i>Acte</i>	<i>Précision</i>	<i>Rappel</i>	<i>F-mesure</i>
opn	0,75	0,675	0,711	conf	0,554	0,486	0,518
cls	0,44	0,52	0,477	expl	0,534	0,476	0,503
gr	0,895	0,64	0,746	justif	0,667	0,597	0,630
intr	0,391	0,25	0,305	part_rej	0	0	0
tkn	0,75	0,591	0,661	rej	0,217	0,37	0,274
apl	0,32	0,26	0,287	part_acc	0	0	0
pf	0,714	0,513	0,597	acc	0,343	0,42	0,38
req	0	0	0	disap	0	0	0
conf_req	0	0	0	ap	0	0	0
expl_req	0,75	0,567	0,646	op	0,42	0,825	0,557
justif_req	1	0,438	0,609	ack	0,718	0,674	0,695
op_req	0,8	0,45	0,576	tm	0,247	0,218	0,232
q	0,458	0,467	0,462	non_und	0	0	0
prm	0	0	0	clarif_req	0	0	0
hop	0	0	0	fback	0,193	0,25	0,218
wsh	0,15	0,22	0,178	intrp	0	0	0
arg	0,378	0,201	0,262	ans	0,1	0,13	0,113
hyp	1	0,559	0,717	stat	0	0	0
clarif	0	0	0	oth	0	0	0
conc	0,862	0,629	0,727	<b>Taux moyen</b>	<b>0,359</b>	<b>0,301</b>	<b>0,318</b>

*Tableau 4.7: Résultats d'évaluation par acte du corpus CARD 1.0*

Nous retenons ainsi les points suivants :

- Les actes de gestion des obligations sociales telle que la présentation des participants (تقديم) présentent des taux faibles vu l'absence de marqueurs lexicaux pour prédire cet acte. En effet, la présentation des différents invités au débat se fait généralement par l'introduction de son nom (معنا من تونس السيدة سهام بن سدرين / avec nous de la Tunisie madame Sihem Ben Sedrine). Ainsi, le modèle de prédiction ne peut pas détecter les entités nommées (de type personne).
- Les actes de demandes (demande / طلب, demande de confirmation / طلب تأكيد et demande de clarification / طلب توضيح) n'ont pas été détectés (valeurs d'évaluation nulles) dans le corpus étant donné qu'ils peuvent se présenter sous plusieurs formes; ce qui complique la tâche de classification utilisant des descripteurs statistiques.
- D'autres actes de demandes (promesse / وعد, espérance / ترحي et souhait / تمنى) ne sont pas correctement classés par le modèle CRF puisqu'ils sont peu fréquents dans le corpus donc leur négligeable nombre d'occurrences engendre une fausse prédiction des actes.
- L'acte argumentatif clarification / توضيح n'est pas détecté par le classificateur CRF vu qu'il est reconnu par le contexte sémantique du débat et il est énoncé suite à une demande de clarification qui n'est pas reconnue par le système.
- Les actes argumentatifs (Appréciation / استحسان et désapprobation / استهجان) sont mal classés et considérés respectivement comme étant des formes d'acceptation et de rejet par le modèle CRF. Ainsi, il est difficile de distinguer entre appréciation/acceptation et désapprobation/rejet.
- L'acceptation partielle / موافقة جزئية et le rejet partiel / رفض جزئي ne sont pas reconnus par le modèle de prédiction. Cela s'explique par la faible occurrence de ces actes dans le corpus. De plus, la reconnaissance de ces actes nécessite des patrons syntaxiques vu que ces actes se présentent généralement sous les formes « oui ... mais » pour l'acceptation partielle et « non... mais » pour le cas du rejet partiel. Ainsi, le classificateur considère l'acceptation partielle comme acceptation et le rejet partiel comme une forme de rejet.
- Le modèle CRF trouve des difficultés à annoter les actes de gestion de dialogue telles que la gestion des tours de parole / إدارة الحوار, le signal de non compréhension / علامة عدم فهم, le dialogue interrompu / حوار متبثر, le retour d'information / تنكير ainsi que les demandes de clarification. Ces actes se basent principalement sur la sémantique dans leur contexte

d'énonciation. Ainsi, il n'y a pas de mots clés spécifiques pour gérer le bon déroulement du débat. D'où, la complexité de la classification automatique de ces actes.

- Les *réponses* / *إجابة* / sont difficiles à détecter vu que le seul indicateur de leur présence est qu'elles sont généralement précédées par des questions.
- Les *déclarations* / *خبر* ne sont pas reconnues par le modèle de classification étant donné qu'il n'existe pas de marqueurs spécifiques pour leur extraction.

Le schéma d'annotation SADA40 utilisé pour l'étiquetage du corpus CARD 1.0 renferme 40 actes répartis sur six catégories à savoir la gestion des obligations sociales, la gestion de dialogue, la demande, l'argumentation, la réponse et la déclaration. Afin de dégager les niveaux de défaillance dans chaque catégorie, nous avons étudié les résultats de la classification des actes de dialogue dans les six groupes appropriés dans le tableau 4.8.

<i>Classifieur</i>	<i>Correct</i>	<i>Incorrect</i>	<i>Précision</i>	<i>Rappel</i>	<i>F-mesure</i>
<i>J48</i>	29,45 %	70,55 %	35,8 %	29,5 %	25,2 %
<i>NB</i>	24,98 %	75,02 %	24,5 %	25 %	24,3 %
<i>SMO</i>	31,7 %	68,3 %	33,8 %	31,7 %	29,8 %
<b><i>CRF</i></b>	<b>32,07 %</b>	<b>67,93 %</b>	<b>35,9 %</b>	<b>32,1 %</b>	<b>31,8 %</b>

*Tableau 4.8: Résultats de la classification du corpus CARD 1.0*

Les résultats d'évaluation de chaque catégorie d'actes de dialogue sont détaillés dans le tableau 4.9. Nous remarquons que les classes *réponse* et *déclaration* ne sont pas correctement identifiées vu le faible nombre de leurs instances ainsi que le manque de critères discriminants pour leur classification.

Classe	Etiquette	Précision	Rappel	F-mesure
إدارة_الالتزامات_الاجتماعية	<i>som</i>	0,656	0,764	0,706
الإنتشاء_الطليبي	<i>req</i>	0,58	0,506	0,540
حجاجي	<i>arg</i>	0,64	0,826	0,721
إدارة_الحوار	<i>tm</i>	0,441	0,34	0,384
إجابة	<i>ans</i>	0,1	0,13	0,113
خبر	<i>stat</i>	0	0	0
<i>Taux moyen</i>		<i>0,356</i>	<i>0,321</i>	<i>0,318</i>

*Tableau 4.9: Résultats de la classification du corpus CARD 1.0 par catégorie*

Les expérimentations montrent que les actes argumentatifs sont les mieux classés (f-mesure=72,1%) mais l'ambiguïté réside dans la reconnaissance de l'acte argumentatif approprié. D'après l'étude de la matrice de confusion, nous remarquons qu'il y a une confusion entre les actes « opinion » et « argument » vu le manque d'indices lexicaux spécifiques à chacun d'eux. En effet, ces actes sont reconnus par le contexte sémantique dans la discussion. De plus,

l'acte « clarification » est généralement confondu avec l'explication étant donné qu'il y a une petite nuance entre les deux. En effet, une clarification peut être considérée comme une forme d'explication.

#### 4.3.1.2. Résultats de CARD 1.1

Afin d'améliorer les résultats obtenus par la première version du corpus de débats CARD 1.0, nous avons réduit le schéma d'annotation initial à 19 actes en effectuant les changements suivants :

- Regroupement des actes de gestion des obligations sociales tels que les formalités d'ouverture et de fermeture du dialogue ainsi que les formules de politesse, les remerciements et les excuses en un seul acte noté « SOM » vu que l'objectif principal est de détecter les actes argumentatifs.
- Regroupement des actes relatifs à la gestion de l'enchaînement des tours de parole dans la discussion (acquiescement, apaisement, etc.) en un seul acte noté « TM » vu la difficulté de distinguer entre les différents actes de cette catégorie.
- Fusion de l'acte « acceptation partielle » avec l'acte « acceptation ».
- Fusion de l'acte « rejet partiel » avec l'acte « rejet ».
- Élimination de certains actes (clarification, déclaration, etc.) vu leur faible fréquence dans le corpus.

Les modifications apportées au schéma d'annotation ont été suivies d'une augmentation de la taille du corpus de débats utilisé afin d'améliorer les résultats de l'apprentissage des actes de dialogue. Ainsi le nouveau schéma noté SADA19 (détaillé dans la section 3.3 du chapitre 2) a été appliqué pour l'annotation de la nouvelle version du corpus (CARD1.1) formée de 22 débats politiques.

Afin de confirmer l'amélioration de la performance des modèles d'apprentissage, nous avons détaillé les résultats d'évaluation en termes de précision, rappel et f-mesure dans le tableau 4.10. Ce tableau montre une étude comparative du modèle CRF avec les classificateurs SMO, J48 et NB.

	Correct	Incorrect	Précision	Rappel	F-mesure
J48 (%)	60,44	39,56	45,3	46,4	44,6
NB (%)	48,02	51,98	48,7	48	45,7
SMO (%)	52,39	47,61	52,6	52,4	51,2
CRF (%)	54,41	45,59	55,9	53,4	54,5

Tableau 4.10: Résultats d'évaluation de la tâche d'annotation

Les résultats obtenus montrent une amélioration remarquable des valeurs d'évaluation pour tous les classificateurs en utilisant un corpus plus important avec un nombre réduit de classes. Ce qui confirme l'importance de la taille du corpus dans le processus d'apprentissage.

Les modèles CRF donnent les meilleurs résultats en termes de précision (55,9%), rappel (53,4%) et f-mesure (54,5%) en intégrant des attributs lexicaux (mots clés et mots de question), morpho-syntaxiques (POS), contextuels (acte précédent et locuteur précédent) et structurels (nom et rôle du locuteur).

Afin de valider l'importance de tous les descripteurs dans la prédiction des actes de dialogue appropriés à chaque énoncé, nous avons effectué plusieurs expérimentations en intégrant à chaque fois un niveau des caractéristiques.

- *Première expérimentation* : inclut uniquement des critères lexicaux.
- *Deuxième expérimentation* : inclut des critères lexicaux et morpho-syntaxiques.
- *Troisième expérimentation* : on intègre des descripteurs niveau énoncé aux caractéristiques de la deuxième expérimentation.
- *Quatrième expérimentation* : inclut tous les descripteurs y compris les critères structurels.

Nous détaillons les résultats d'apprentissage des diverses expérimentations menées, basées sur les différents modèles de prédiction dans les tableaux 4.11, 4.12 et 4.13 et 4.14.

Descripteurs	Précision	Rappel	F-mesure
Expérimentation 1	0,50	0,507	0,491
Expérimentation 2	0,503	0,51	0,493
Expérimentation 3	0,511	0,519	0,503
<b>Expérimentation 4</b>	<b>0,526</b>	<b>0,524</b>	<b>0,512</b>

*Tableau 4.11: Résultats d'annotation du classifieur SMO*

Descripteurs	Précision	Rappel	F-mesure
Expérimentation 1	0,483	44,9	0,426
Expérimentation 2	0,476	0,45	0,427
Expérimentation 3	0,474	0,47	0,442
<b>Expérimentation 4</b>	<b>0,48</b>	<b>0,487</b>	<b>0,457</b>

*Tableau 4.12: Résultats d'annotation du classifieur NB*

Descripteurs	Précision	Rappel	F-mesure
Expérimentation 1	0,401	0,423	0,372
Expérimentation 2	0,425	0,441	0,388
Expérimentation 3	0,44	0,437	0,433
Expérimentation 4	0,453	0,464	0,446

*Tableau 4.13: Résultats d'annotation du classifieur J48*

Descripteurs	Précision	Rappel	F-mesure
Expérimentation 1	0,51	0,528	0,516
Expérimentation 2	0,503	0,51	0,493
Expérimentation 3	0,511	0,519	0,503
<b>Expérimentation 4</b>	<b>0,559</b>	<b>0,534</b>	<b>0,545</b>

Tableau 4.14: Résultats d'annotation du classifieur CRF

Les différentes expérimentations montrent l'importance des descripteurs lexicaux qui ont permis la reconnaissance de plus de 50% des actes en appliquant les modèles CRF (F-mesure=51,6%). La combinaison de tous les critères apporte une amélioration de 2% à 7% dans les résultats d'évaluation des différents classifieurs.

Afin d'analyser les erreurs d'annotation des actes de dialogue, nous avons effectué une étude détaillée des résultats d'évaluation de chaque acte à étiqueter dans le corpus CARD 1.1. Les résultats de la classification obtenus par les modèles CRF sont illustrés dans le tableau 4.15.

Acte de dialogue	Etiquette	Précision	Rappel	F-mesure
<b>Gestion des obligations sociales</b>	<b>Som</b>	<b>0,814</b>	<b>0,787</b>	<b>0,800</b>
Gestion de dialogue	Tm	0,421	0,492	0,454
<b>Question</b>	<b>Q</b>	<b>0,71</b>	<b>0,692</b>	<b>0,701</b>
Demande de confirmation	conf_req	0,386	0,376	0,381
<b>Demande de justification</b>	<b>justif_req</b>	<b>0,72</b>	<b>0,716</b>	<b>0,718</b>
Demande d'opinion	op_req	0,318	0,378	0,345
Demande d'explication	expl_req	0,644	0,726	0,683
Opinion	Op	0,553	0,685	0,612
Thèse	Thesis	0,362	0,367	0,364
Argument	Arg	0,375	0,222	0,279
Confirmation	Conf	0,498	0,42	0,456
Hypothèse	Hyp	0,69	0,67	0,680
<b>Conclusion</b>	<b>Conc</b>	<b>0,875</b>	<b>0,656</b>	<b>0,750</b>
Explication	Expl	0,639	0,614	0,626
<b>Justification</b>	<b>Justif</b>	<b>0,879</b>	<b>0,724</b>	<b>0,794</b>
Rejet	Rej	0,58	0,557	0,568
Acceptation	Acc	0,605	0,615	0,610
Réponse	Ans	0,36	0,296	0,325
Autres	Oth	0,193	0,159	0,174
<b>Taux Moyen</b>		<b>0,559</b>	<b>0,534</b>	<b>0,545</b>

Tableau 4.15: Résultats d'évaluation du classifieur CRF avec CARD 1.1

D'après l'étude détaillée des résultats d'annotation de chaque acte, nous pouvons déduire les points suivants:

- L'acte "gestion des obligations sociales" (som) donne les meilleurs résultats (f-mesure =80%), ce qui prouve que le regroupement des actes formant cette catégorie permet une meilleure

reconnaissance avec les mots clés spécifiques tels que أهلا /salut, مرحبا /bienvenue, عفوا /excuse-moi, شكرًا / merci et إلى اللقاء / au revoir.

- L'acte question (*q*) est reconnu à 70% par les mots de question considérés comme critère discriminant pour l'identification de tout type de question.
- Les actes conclusion (*conc*) et justification (*justif*) montrent de bonnes valeurs d'évaluation (f-mesure=75% pour l'acte conclusion et f-mesure=79,4% pour l'acte justification). En effet, la catégorie conclusion est généralement reconnue par les expressions إذن /donc, بالتالي /en conclusion et نستنتج /on conclut. Alors que pour exprimer une justification on utilise souvent le marqueur lexical لأن / parce que.
- Les actes exprimant une demande de justification (*justif\_req*) sont relativement reconnus par le modèle d'apprentissage (f-mesure=72,8%) grâce au mot de question "لماذا / pourquoi" utilisé souvent pour demander une justification. De même, les actes de demande d'explication (*expl\_req*) débutant généralement par le mot de question "كيف / comment" sont souvent étiquetés correctement par le modèle CRF (f-mesure=68,3%).

#### 4.3.1.3. Résultats de CARD 1.2

Pour une meilleure reconnaissance des actes de dialogue, nous avons amélioré la taille du corpus d'apprentissage en annotant 50 débats contenant 14062 actes et 260212 mots.

La nouvelle version du corpus (notée CARD 1.2) a été annotée en utilisant le schéma réduit SADA19 qui inclut 19 catégories d'actes. Les résultats d'évaluation des différents modèles de classification appliqués pour l'identification des actes de dialogue du corpus CARD 1.2 sont présentés dans le tableau 4.16.

	Précision	Rappel	F-mesure
CRF (%)	64,7	62,2	63,43
SMO (%)	59,9	58,5	58,19
N B (%)	56,6	54,2	53,37
J48 (%)	52,9	51,3	50,09

Tableau 4.16: Résultats d'évaluation du corpus CARD 1.2

Ces résultats justifient le choix du classificateur CRF vu qu'il donne une meilleure classification avec une valeur de f-mesure égale à 63,43%. De plus, nous pouvons déduire l'importance de la taille du corpus dans la tâche d'annotation des actes de dialogue. En effet, nous avons obtenu

une amélioration d'environ 9% en augmentant le nombre d'actes de 6050 (taille du corpus CARD 1.1) à 14062 actes dans la dernière version du corpus CARD.

Afin d'analyser les erreurs d'annotation des actes de dialogue, nous avons effectué une étude détaillée des résultats d'évaluation de chaque acte à étiqueter dans le corpus CARD 1.2. Les résultats de la classification obtenus par les modèles CRF sont illustrés dans le tableau 4.17.

D'après l'étude détaillée des résultats d'annotation de chaque acte, nous pouvons déduire les points suivants:

<i>Acte de dialogue</i>	<i>Etiquette</i>	<i>Précision</i>	<i>Rappel</i>	<i>F-mesure</i>
<b><i>Gestion des obligations sociales</i></b>	<b><i>Som</i></b>	<b><i>0,894</i></b>	<b><i>0,867</i></b>	<b><i>0,880</i></b>
<i>Gestion de dialogue</i>	<i>Tm</i>	<i>0,660</i>	<i>0,644</i>	<i>0,652</i>
<b><i>Question</i></b>	<b><i>Q</i></b>	<b><i>0,810</i></b>	<b><i>0,792</i></b>	<b><i>0,801</i></b>
<i>Demande de confirmation</i>	<i>conf_req</i>	<i>0,516</i>	<i>0,496</i>	<i>0,506</i>
<i>Demande de justification</i>	<i>justif_req</i>	<i>0,790</i>	<i>0,780</i>	<i>0,785</i>
<i>Demande d'opinion</i>	<i>op_req</i>	<i>0,598</i>	<i>0,571</i>	<i>0,584</i>
<i>Demande d'explication</i>	<i>expl_req</i>	<i>0,794</i>	<i>0,782</i>	<i>0,788</i>
<i>Opinion</i>	<i>Op</i>	<i>0,593</i>	<i>0,605</i>	<i>0,599</i>
<i>Thèse</i>	<i>Thesis</i>	<i>0,422</i>	<i>0,386</i>	<i>0,403</i>
<i>Argument</i>	<i>Arg</i>	<i>0,395</i>	<i>0,340</i>	<i>0,365</i>
<i>Confirmation</i>	<i>Conf</i>	<i>0,590</i>	<i>0,560</i>	<i>0,575</i>
<i>Hypothèse</i>	<i>Hyp</i>	<i>0,740</i>	<i>0,790</i>	<i>0,764</i>
<b><i>Conclusion</i></b>	<b><i>Conc</i></b>	<b><i>0,894</i></b>	<b><i>0,866</i></b>	<b><i>0,880</i></b>
<i>Explication</i>	<i>Expl</i>	<i>0,690</i>	<i>0,663</i>	<i>0,676</i>
<b><i>Justification</i></b>	<b><i>justif</i></b>	<b><i>0,890</i></b>	<b><i>0,860</i></b>	<b><i>0,875</i></b>
<i>Rejet</i>	<i>Rej</i>	<i>0,682</i>	<i>0,620</i>	<i>0,650</i>
<i>Acceptation</i>	<i>Acc</i>	<i>0,690</i>	<i>0,625</i>	<i>0,656</i>
<i>Réponse</i>	<i>Ans</i>	<i>0,390</i>	<i>0,336</i>	<i>0,361</i>
<i>Autres</i>	<i>Oth</i>	<i>0,250</i>	<i>0,229</i>	<i>0,239</i>
		<b><i>0,647</i></b>	<b><i>0,622</i></b>	<b><i>0,643</i></b>

**Tableau 4.17:** Résultats d'évaluation du classifieur CRF avec CARD 1.2

- L'acte argumentatif *thèse* n'est pas bien reconnu par le modèle CRF (f-mesure=40,3%). Ce faible score est dû à la difficulté de détecter les idées énoncées par le locuteur. Une thèse exprime généralement une problématique à discuter par les participants au débat qui sera par la suite retenue ou rejetée. Cet acte est souvent introduit par des phrases déclaratives ne contenant pas des indices lexicaux spécifiques. Ainsi la reconnaissance automatique de cet acte se base essentiellement sur le contexte d'énonciation et sur la sémantique du discours. Donc les critères utilisés par le modèle sont insuffisants pour la détection de l'étiquette correcte lors du processus d'annotation automatique. Le tableau 4.18 illustre des exemples de l'acte thèse.

<p>الأستاذ المولدي لحمر كلما تقدم الحوار الوطني في تونس خطوة أو خطوة إلا وجاءت ضربة موجعة كسرت ظهر هذا الحوار.</p> <p><i>Monsieur Mouloudi Lahmar, au fur et à mesure que progressait le dialogue national en Tunisie, un coup douloureux a brisé le dos de ce dialogue.</i></p>
<p>الكثيرون يختزلون حاجة تونس أساسا إلى علاقات قوية مع الجزائر يختزلونها في الملف الأمني خاصة بعد تفشي ظاهرة الإرهاب على الحدود.</p> <p><i>Beaucoup réduisent la nécessité pour la Tunisie d'entretenir de solides relations avec l'Algérie, notamment dans le dossier de la sécurité, en particulier après la propagation du terrorisme à la frontière.</i></p>
<p>أشار رئيس الحكومة رجب طيب أر دوغان إلى وجود أيدي خارجية تعبت بالوضع في تركيا</p> <p><i>Le Premier ministre Rajeb Tayyip Er Dogan a souligné la présence de mains extérieures altérant la situation en Turquie.</i></p>

Tableau 4.18: Exemples de l'acte de dialogue "Thèse"

- L'acte exprimant un argument donne de faibles résultats (f-mesure = 36,5%). Ainsi il est difficile de détecter les actes argumentatifs surtout en absence de marqueurs lexicaux facilitant la reconnaissance. Cet acte est souvent confus avec l'acte exprimant une opinion. L'exemple présenté dans le tableau 4.19 illustre une fausse annotation de l'acte *argument* et sa confusion avec l'acte *opinion*.

Enoncé	Acte prédit	Acte correct
<p>بداية الاعتقاد بشرعية قوى سياسية معينة كانت محرومة حتى من التعبير عن نفسها، هذا أيضا إنجاز و إنجاز مهم القبول بأن يصبح الشارع له دوره في اتخاذ القرار.</p> <p><i>Le début de la croyance en la légitimité de certaines forces politiques a même été privé de toute expression, ce qui constitue également une réalisation importante et une acceptation du fait que la rue a un rôle à jouer dans la prise de décision.</i></p>	Opinion	Argument

Tableau 4.19: Exemple d'une fausse annotation de l'acte de dialogue "Argument"

La difficulté de reconnaissance de l'acte argumentatif est expliquée par la variété et la complexité des types d'arguments qui peuvent occurer dans une discussion politique. Parmi les types d'arguments détectés, nous citons :

- *Argument de la réalité/ حجة الواقع* : le participant au débat essaie de convaincre les autres locuteurs de son point de vue en citant des faits ou des événements issus de la vie réelle. Des exemples d'arguments de réalité sont présentés dans le tableau 4.20.

ونحن نعلم في الستينات كيف كانت محطة عربية تدير الناس في الشوارع Nous savons dans les années 1960 comment une station arabe dirigeait les gens dans la rue
وأقرب مثل شاهدناه في ما حدث ما بين مصر و الجزائر في الرياضة Le plus proche que nous avons vu est ce qui s'est passé entre l'Égypte et l'Algérie dans le sport
في السودان في الفترة الماضية كان هناك تشجيع كبير من بعض الصحفيين Au Soudan, au cours de la période écoulée, certains journalistes ont été très encouragés

Tableau 4.20: Exemples d'arguments de la réalité

- *Argument de comparaison/ حجة المماثلة* : le locuteur met l'accent sur les points communs entre deux situations similaires pour persuader les autres participants par son opinion. Le tableau 4.21 détaille des exemples d'arguments de comparaison.

حتى في ثورات كبرى حتى في تاريخ الثورة الفرنسية حتى في إيران عندما حصل قبل ثلاثين عاما أشهر تقريبا Même dans les grandes révolutions, même dans l'histoire de la Révolution française, même en Iran quand cela s'est passé il y a presque trente ans
الحرية لها ثمن و الكرامة لها ثمن La liberté a un prix et la dignité a un prix.

Tableau 4.21: Exemples d'arguments comparaison

- *Argument d'autorité/ حجة الشاهد القولي* : le locuteur s'appuie principalement sur une citation pour renforcer son point de vue sur une idée donnée. Ce type d'argument est présenté dans le tableau 4.22.

في جملة شهيرة الجميع يرددونها، الشعب يريد إسقاط النظام Dans une phrase célèbre que tout le monde répète, le peuple veut renverser le régime
كما قال أبو القاسم الشابي إذا الشعب يوما أراد الحياة Comme a déclaré Abu al-Qasim Chebbi que si les gens voulaient un jour vivre

Tableau 4.22: Exemples d'arguments d'autorité

- *Argument logique/ حجة منطقية* : la persuasion se fait à travers un argument rationnel basé sur la l'adhésion intellectuelle et la logique des faits. L'exemple suivant illustre ce type d'argument:

**النظرية تقول بأنه ينبغي على الصحفي أن يؤمن بالمثل العليا للصحافة**

*La théorie est qu'un journaliste devrait croire aux idéaux du journalisme et de l'éthique des médias.*

#### 4.3.2. Analyse des cas d'erreurs

L'échec d'identification de des actes de dialogue peut s'expliquer par la présence de quelques erreurs. Nous analysons ces erreurs et nous résumons les causes dans les points suivants :

- *Erreurs dues à l'analyse morphologique*

Dans la langue arabe, plusieurs mots sont ambigus, c'est-à-dire, qu'ils peuvent avoir plus qu'une catégorie grammaticale. Par exemple, le mot أحمد /OHmd/Ahmed peut être un verbe أحمَدُ /OaHomadu/ je remercie ou un nom propre أَحْمَدُ /OaHomado/Ahmed.

- *Erreurs dues à l'ambiguïté des mots clés*

Certains mots clés peuvent causer des erreurs dans l'identification des actes de dialogue. En effet le mot « يعني » généralement employé pour présenter une explication peut être utilisé comme une expression triviale dans l'oral spontané.

La non voyelation des mots clés peut aussi causer des ambiguïtés dans la reconnaissance des actes appropriés. Par exemple le mot « إذا » peut exprimer une hypothèse (إذا/si) comme il peut indiquer être utilisé pour conclure une idée (إِذَا/donc).

L'ambiguïté dans l'identification des actes de dialogue peut être due aussi suite à la présence de deux mots clés successifs exprimant deux actes différents.

Par exemple l'acte(A1)peut provoquer une fausse annotation (acceptation) vu la présence du mot clé "نعم" au début de l'acte alors qu'il s'agit d'un acte présentant une opinion illustrée par le mot clé "أعتقد".

(A1) "نعم أعتقد أن عمليات بناء مؤسسات النظام الجديد تتسم بالبطء الشديد"

- *Erreurs dues à l'absence de marqueurs lexicaux*

Certains actes argumentatifs sont reconnus par le contexte sémantique de leur énonciation. Il est ainsi difficile de les identifier avec les caractéristiques d'apprentissage basées principalement sur des indices lexicaux. Dans ce contexte, il est difficile de distinguer entre une opinion et un argument, une explication et une opinion, une expression de gestion de dialogue et un acte argumentatif. D'où la complexité de la tâche d'annotation surtout des actes argumentatifs.

## 5. Conclusion

Ce chapitre englobe une série d'expérimentations effectuées pour les deux principaux modules du système RECAARD (**RE**Cognition of **Act** in **Ar**abic **De**bates) à savoir les modules de segmentation et d'annotation des actes de dialogue. Nous avons présenté les résultats d'évaluation obtenus en appliquant les modèles CRF. Ainsi, nous avons comparé les résultats obtenus avec ceux des classificateurs SMO, Naïve Bayes et J48 sur trois versions différentes du corpus de débats politiques « CARD ».

Les résultats obtenus sont encourageants et justifient l'apport apporté par les modèles CRF dans les tâches de segmentation et d'annotation des actes de dialogue.

Nous pouvons aussi confirmer l'importance de la taille du corpus dans l'amélioration des résultats de classification des actes de dialogue.

## CONCLUSION ET PERSPECTIVES

Les travaux menés dans cette thèse s'intègrent dans le cadre de l'analyse des discussions Homme-Homme et s'intéressent plus particulièrement au problème de la reconnaissance des actes de dialogue. Cette problématique a été abordée pour la langue arabe dans un contexte cerné à savoir les débats politiques télévisés. Ces derniers constituent une ressource riche en actes argumentatifs et reflètent la dynamique du dialogue à travers l'exposition des attitudes et des opinions des intervenants.

Afin d'aborder cette problématique, nous avons commencé par étudier l'état de l'art réalisé dans le cadre des tâches constituant le processus de reconnaissance des actes de dialogue à savoir la segmentation et l'annotation. Nous avons constaté que la plupart des recherches se sont basées sur l'apprentissage automatique pour identifier les actes de dialogue en s'appuyant sur des corpus de dialogues spécifiques au contexte du travail.

Ainsi, l'élaboration du corpus CARD (**C**orpus of **AR**abic **D**ebates) a été le point de départ de cette thèse. Nous avons pu collecter des transcriptions textuelles des débats télévisés traitant des thématiques politiques d'actualité telles que la révolution et les élections tunisiennes. Ensuite et moyennant ce corpus, nous avons proposé un schéma d'annotation (SADA: **S**cheme **A**notation for **D**ebates in **A**rabic) spécifique au contexte argumentatif des débats et basé sur la taxonomie générale de la langue arabe. Ce schéma a été réalisé en une première étape avec une taxonomie contenant 40 actes de dialogue (SADA40). Cette dernière a été ensuite réduite à 19 actes de dialogue (SADA19) afin d'améliorer les résultats de l'apprentissage. Les deux versions du schéma SADA ont été validées par des experts. Le corpus de débats a été réalisé via l'outil ACTAAR (**ACT**Anotation in **AR**abic) qui permet l'étiquetage semi-automatique des actes de dialogue en utilisant le schéma SADA. Ainsi, nous avons pu construire un corpus contenant 260383 mots et 14062 actes de dialogue à partir de 50 débats télévisés.

Nous avons ensuite proposé une méthode de reconnaissance d'actes de dialogue qui repose sur les modèles CRF (Conditional Random Fields) que nous avons appliqués sur les deux étapes de segmentation et d'annotation des actes de dialogue. Le choix des champs aléatoires conditionnels a été renforcé par leur performance dans les tâches d'étiquetage des données séquentielles. De plus, ces modèles prennent en considération les corrélations et les dépendances qui coexistent entre les mots d'un même énoncé.

L'originalité de notre méthode de reconnaissance des actes de dialogue réside dans l'annotation des discussions argumentatives permettant d'explicitier la dynamique du dialogue en termes de conflits, d'opinions divergentes, d'explications et d'arguments soutenus ou rejetés par les différents intervenants tout au long du débat.

La construction du corpus de débats constitue aussi un apport majeur dans cette thèse étant donné le problème de manque de ressources pour la langue arabe. Notre corpus constitue ainsi une ressource importante validée par des experts humains qui peut être exploitée dans les travaux portant sur les dialogues Homme-Homme.

L'implémentation de la méthode de reconnaissance proposée a donné naissance au système RECAARD (**RE**ognition of **A**cts in **AR**abic **D**ebates) constitué principalement de trois modules à savoir le prétraitement, la segmentation et l'annotation des débats télévisés.

Les résultats d'évaluation de la méthode proposée ont mis en relief les performances des modèles CRF et ont montré expérimentalement leur apport en comparaison avec d'autres classificateurs à savoir les réseaux bayésiens naïfs (Naive Bayes), les machines à vaste marge (SMO) et les arbres de décision (J48). En effet, nous avons obtenu des valeurs de F-mesure égale à 97,9% pour la tâche de segmentation en énoncés et 63,4% pour l'annotation des énoncés segmentés en actes de dialogue.

Ces résultats sont très encourageants et ouvrent la voie à de nouvelles perspectives que nous résumons dans les points suivants.

Comme première perspective à court terme, nous envisageons de réaliser une évaluation extrinsèque de notre système de reconnaissance des actes de dialogue dans une application du TALN telle qu'un résumeur automatique des discussions ou un système de Question/réponse spécifique aux questions d'opinions afin de mesurer l'apport de la tâche d'identification des actes de dialogue dans ces applications.

Par ailleurs, étant donnée la nature argumentative des débats télévisés, nous visons aussi à court terme étendre notre approche de reconnaissance des actes de dialogue pour générer les structures argumentatives reflétant l'enchaînement des actes tout au long de la discussion. Ce qui consiste à regrouper les actes de dialogue afin de former des séquences d'actes. Le point de départ de l'extraction des séquences d'actes de dialogue est la construction des paires adjacentes telles que question/réponse, opinion/refus, demande de confirmation/confirmation, etc. Ces paires facilitent la génération d'une structure argumentative de discussions qui peut aider l'utilisateur à répondre à des questions complexes telle que « Qui a rejeté l'opinion de X ? ».

A long terme et afin d'améliorer les résultats d'apprentissage du processus de reconnaissance des actes de dialogue, nous visons tenir compte des aspects suivants:

- Élargir la taille du corpus en annotant de nouvelles discussions par des experts humains.
- Intégrer des critères d'ordres sémantique et pragmatique dans la tâche d'annotation afin de pallier les difficultés d'identification des actes argumentatifs tels que les opinions et les arguments qui manquent souvent d'indices lexicaux.
- Intégrer des critères prosodiques qui expriment la sémantique émotionnelle des différents intervenants dans le débat telles que la durée de la pause entre les mots, les mesures de la hauteur et de l'énergie de la voix qui peuvent donner une idée sur l'intention de l'intervenant en énonçant un acte de dialogue.
- Combiner les modèles CRF avec d'autres méthodes d'apprentissage tels que les réseaux de neurones récurrents qui ont montré récemment leur performance dans le domaine de reconnaissance des actes de dialogue.

## BIBLIOGRAPHIE

- (Allen et Core, 1997) Allen, J. and M. Core. 1997. "Draft of DAMSL: Dialog Act Markup in Several Layers". Technical report.
- (Anderson et al., 1991) Anderson, A., M. Bader, E. Bard, E. Boyle, and G. Doherty. 1991. "TheHCRC Map Task Corpus". *Language and Speech*, 34:351-366.
- (Ang et al., 2005) Ang, J., Y. Liu, and E. Shriberg. 2005. "Automatic Dialog Act Segmentation and Classification in Multiparty Meetings". In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1061-1064, Philadelphia.
- (Austin, 1975) Austin, J. L. 1975. "How to do things with words". *Oxford University Press*, 619-650.
- (Bahou et al., 2008) Bahou Y., Belguith L. H and BenHamadou A. 2008. "Towards a Human-Machine Spoken Dialogue in Arabic," In *Proceeding of Workshop on HLT & NLP within the Arabic world: Arabic Language and local languages processing: Status Updates and Prospects*, at the 6th Language Resources and Evaluation Conference (LREC08), Marrakech, Maroc.
- (BenDbabis et al., 2015) BenDbabis S., H. Ghorbel, L. Belguith and M. Kallel. 2015. "Automatic dialogue acts Annotation within Arabic debates". *16th International Conference on Intelligent Text Processing and Computational Linguistics* , April 14-20, Cairo, Egypt.
- (BenDbabis et al., 2016) BenDbabis S., B. Rguii, H. Ghorbel and L. Belguith. 2016. "Utterance Segmentation Using Conditional Random Fields". *27th International Business on Information Management Association*, May 3-5, 2016, Milano, Italy.
- (BenDbabis et al., 2012) BenDbabis S., F. Mallek, H. Ghorbel and L. Belguith. 2012. "Dialogue Acts Annotation Scheme within Arabic discussions". *Sixteenth Workshop on the Semantics and Pragmatics of Dialogue*, September 19-21, Paris, France.
- (BenDbabis et al., 2018) BenDbabis S., H. Ghorbel, L. Belguith. 2018. "Sequential dialogue act recognition for Arabic argumentative debates". *Procesamiento del Lenguaje Natural (SEPLN)* ,60, 53-60., 2018.
- (Blandón et al., 2019) Blandón, M. A. C., Minnema, G., Nourbakhsh, A., Boritchev, M., & Amblard, M. (2019). "Toward dialogue modeling: A semantic annotation scheme for questions and answers. In *Proceedings of the 13th Linguistic Annotation Workshop* (230-235).
- (Blasiak et al., 2011) Blasiak, S.; Rangwala, H. 2011. "A Hidden Markov Model Variant for Sequence Classification". *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*. 22: 1192.
- (Bothe et al., 2018) Bothe, C., Weber, C., Magg, S., & Wermter, S. 2018. "A Context-based Approach for Dialogue Act Recognition using Simple Recurrent Neural Networks". In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.
- (Brill, 1995) Brill, E. 1995. "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging". *Computational Linguistics*, 21, 543-565.
- (Bunt, 2009) Bunt H. 2009. "The DIT++ taxonomy for functional dialogue markup". In *Proceedings of the AAMAS Workshop*, Budapest, May 12, 2009.
- (Carletta et al., 2005) Carletta J., S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D.

- Reidsma, and P. Wellner. 2005. "The AMI Meetings Corpus". In *Proceedings of the Measuring Behavior Symposium on "Annotating and measuring Meeting Behavior"*.
- (Caspers, 2006) Caspers, J. 2006. "Pitch Accents, Boundary Tones and Turn-Taking in Dutch Map Task Dialogues". In *International Conference on Speech and Language Processing*, pages 565-568.
- (Chen et al., 2018) Chen, Z., Yang, R., Zhao, Z., Cai, D., & He, X. 2018. "Dialogue act recognition via crf-attentive structured network". In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 225-234.
- (Clark et al., 2018) Clark, K., Luong, M., Manning, C.D., & Le, Q.V. 2018. "Semi-Supervised Sequence Modeling with Cross-View Training". *EMNLP*.
- (Cohen, 1960) Cohen, J. 1960. "A coefficient of agreement for nominal scales". *Educational and Psychological Measurement*, 20(1):37-46.
- (Core et Allen) Core M. and J. Allen. 1997. "Coding Dialogs with the DAMSL annotation scheme". In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, MIT, Cambridge, MA.
- (Dalva et al., 2018) Dalva, D., Guz, U., & Gurkan, H. 2018. "Effective semi-supervised learning strategies for automatic sentence segmentation". *Pattern Recognition Letters*, 105, 76-86.
- (Diab et al., 2007) Diab, M., Hacioglu, K., and Jurafsky, D. 2007. "Automated Methods for Processing Arabic Text: From Tokenization to Base Phrase Chunking". In *van den Bosch, A. and Soudi, A., editors, Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Kluwer/Springer
- (Di Eugenio et al., 2014) Di Eugenio B., Z. Xie and R.o Serafin. 2014. "Dialogue act classification, higher order dialogue structure, and instance-based learning". *Dialogue and Discourse*, 1(2):1–24.
- (Elmadany et al., 2015) Elmadany A., S. M. Abdou and M. Gheith. 2015. "Turn Segmentation Into Utterances For Arabic Spontaneous Dialogues And Instant Messages". *International Journal on Natural Language Computing*, 4(2). April 2015.
- (Elmadany et al., 2014) Elmadany A., S. M. Abdou, and M. Gheith. 2014. "Arabic Inquiry-Answer Dialogue Acts Annotation Schema". *IOSR Journal of Engineering*, 4(12-V2):32-36.
- (Elmadany et al. 2018) Elmadany, A., Abdou, S., & Gheith, M. 2018. "Improving Dialogue Act classification for Spontaneous Arabic Speech and Instant Messages at Utterance Level". In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.
- (Fomey et David, 2005) Forney Jr, G. David. 2005. "The viterbi algorithm: A personal history." *arXiv preprint cs/0504020*.
- (Geertzen et al., 2004) Geertzen, J., Y. Girard, and R. Morante. 2004. The DIAMOND project. In *the 8th Workshop on Semantics & Pragmatics of Dialogue (CATALOG2004)*, Barcelona, Spain.
- (Geertzen et al., 2007) Geertzen J., V. Petukhova and H. Bunt. 2007. "A Multidimensional Approach to Utterance Segmentation and Dialogue Act Classification". In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp.
- (Geertzen et Bunt, 2006) Geertzen, J. and H. Bunt. 2006. "Measuring Annotator Agreement in a Complex Hierarchical Dialogue Act Annotation Scheme". In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, Association for Computational Linguistics*. 126-133, Sydney, Australia.
- (Georgila et al., 2009) Georgila, K., O. Lemon, J. Henderson, and J. D. Moore. 2009. "Automatic Annotation of Context and Speech Acts for Dialogue Corpora". *Journal of Natural Language Engineering*, 15(3):315-353.
- (Graja et al., 2013) Graja M., Jaoua M. and Belguith L. H. 2013, "Discriminative Framework for Spoken Tunisian Dialect Understanding," In proceeding of the 2nd International Conference on Statistical Language and Speech Processing, SLSP.
- (Grau et al., 2004) Grau S., Sanchis E., Castro M. J., Vilar D. 2004. "Dialogue act classification using a Bayesian approach", In *9th Conference Speech and Computer (SPECOM-2004)*, 495-499.

- (Habash et al., 2013) Habash, N., Roth, R., Rambow, O., Eskander, R., and Tomeh, N. 2013. "Morphological Analysis and Disambiguation for Dialectal Arabic". In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, GA.
- (Hardy et al., 2005) Hardy, H., A. Biermann, R. Bryce Inouye, A. McKenzie, T. Strzalkowski, C. Ursu, N. Webb, and M. Wu. 2005. "The AMITIES System: Data-Driven Techniques for Automated Dialogue". *Speech Communication*, 48:354-373.
- (Hijawi et al., 2013) Hijawi M., Bandar Z., and Crockett K. 2013. "User's Utterance Classification Using Machine Learning for Arabic Conversational Agents," In *Proceeding of 5th International Conference on Computer Science and Information Technology (CSIT)*, pp. 223-232.
- (Ivanovic, 2005a) Ivanovic, E. 2005. "Dialogue act tagging for instant messaging chat sessions". In *Proceedings of the ACL Student Research Workshop. Association for Computational Linguistics*. 79-84.
- (Ivanovic, 2005b) Ivanovic E. 2005. Automatic Utterance Segmentation in Instant Messaging Dialogue. *Proceedings of the Australasian Language Technology Workshop*. Pages 241–249, Sydney, Australia.
- (Ji et Bilmes, 2005) Ji, G., Bilmes, J. 2005. "Dialog act tagging using graphical models". In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*. 1-33.
- (Ji et al., 2006) Ji, Q., Lan, P., & Looney, C. 2006. "A probabilistic framework for modeling and real-time monitoring human fatigue". *IEEE Transactions on systems, man, and cybernetics: Systems and humans*, 36(5), 862-875.
- (Jurafsky et al., 1998) Jurafsky, D., R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor, and C. Van Ess-Dykema. 1998. "Switchboard Discourse Language Modeling Project Final Report". *Research Note 30, Center for Language and Speech Processing*, Johns Hopkins University, Baltimore.
- (Khoufi et al., 2014) Khoufi N., Aloulou C. and Belguith L. H. 2014. "Chunking Arabic Texts Using Conditional RandomFields", In *the Proceedings of The 11th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA 2014)*, November 2014, Doha, Qatar, pp 428432.
- (Kumar et al., 2018) Kumar, H., Agarwal, A., Dasgupta, R., & Joshi, S. 2018. "Dialogue act sequence labeling using hierarchical encoder with crf". In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- (Landis et Koch, 1977) Landis, J. R., & Koch, G. G. 1977. "An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers". *Biometrics*, 363-374.
- (Lendvai et Geertzen, 2007) Lendvai, P. and J. Geertzen. 2007. "Token-Based Chunking of Turn-Internal Dialogue Act Sequences". In *Proceedings of the 8th SIGDIAL Workshop on Discourse and Dialogue*, 174-181, Antwerp, Belgium.
- (Levin et al., 2003) Levin, L., C. Langley, A. Lavie, D. Gates, and D. Wallace. 2003. "Domain Specific Speech Acts for Spoken Language Translation". In *Proceedings of 4th SIGdial Workshop on Discourse and Dialogue*.
- (Lhioui et al., 2013) Lhioui C., Zouaghi A. and Zrigui M. 2013. "A Combined Method Based on Stochastic and Linguistic Paradigm for the Understanding of Arabic Spontaneous Utterances," In *Proceeding of CICLing 2013, Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science*, Samos, Greece, 549-558.
- (Liu et al., 2017) Liu, Y., Han, K., Tan, Z., & Lei, Y. 2017. "Using context information for dialog act classification in DNN framework". In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2170-2178.
- (Maingueneau, 2016) Maingueneau, D. 2016. "Énonciation et analyse du discours". *Corela. Cognition, représentation, langage*, (HS-19).
- (Martínek et al., 2019) Martínek, J., Král, P., Lenc, L., & Cerisara, C. 2019. "Multi-lingual Dialogue Act Recognition with Deep Learning Methods". *arXiv preprint arXiv:1904.05606*.
- (Meteer et al., 1995) Meteer, M. W., Taylor, A. A., MacIntyre, R., & Iyer, R. 1995. "Dysfluency annotation stylebook for the switchboard corpus". Philadelphia, PA: University of Pennsylvania.
- (Nagata et Morimoto, 1994) Nagata, M. and T. Morimoto. 1994. "First Steps Towards Statistical Modeling of Dialogue to Predict the Speech Act Type of the Next Utterance". *Speech Communication*.

- (Pasha et al., 2014) Pasha, A., Al-Badrashiny, M., Diab, M. T., El Kholy, A., Eskander, R., Habash, N., ... & Roth, R. 2014. "Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic". In *LREC*. 14, 1094-1101.
- (Pearl et Russel, 2002) Pearl J, Russell S. 2002. "Bayesian Networks". In Arbib MA (ed.). "Handbook of Brain Theory and Neural Networks". Cambridge, Massachusetts: Bradford Books (MIT Press). 157–160. .
- (Pleva et al., 2015) Pleva, M., Ondáš, S., & Juhár, J. 2015. "Automatic dialogue acts classification in Slovak dialogues". In *2015 25th International Conference Radioelektronika (RADIOELEKTRONIKA)* (pp. 295-298). IEEE.
- (Prasad et Walker, 2002) Prasad, R. and M.Walker. 2002. "Training a Dialogue Act Tagger for Human-Human and Human-Computer Travel Dialogues". In *Proceedings of the 3rd SIGdial workshop on Discourse and Dialogue*, Philadelphia, Pennsylvania.
- (Raheja et Tetreault, 2019) Raheja, V., & Tetreault, J. 2019. "Dialogue Act Classification with Context-Aware Self-Attention". *arXiv preprint arXiv:1904.02594*.
- (Reithinger et Klesen, 1997) Reithinger, N. and M. Klesen. 1997. "Dialogue Act Classification Using Language Models". In *Proceedings of EuroSpeech-97*.
- (Rotaru, 2002) Rotaru, M. 2002. "Dialog Act Tagging using Memory-Based Learning". Termproject, University of Pittsburgh.
- (Samuel et al., 1999) Samuel, K., S. Carberry, and K. Vijay-Shanker. 1999. "Automatically Selecting Useful Phrases for Dialogue Act Tagging". In *Proceedings of the Fourth Conference of the Pacific Association for Computational Linguistics*, Waterloo, Ontario, Canada.
- (Searle, 1968) Searle, J. R. 1968. "Austin on locutionary and illocutionary acts". *The philosophical review*, 77(4), 405-424.
- (Shala et al., 2010) Shala L., Rus V. and Graesser A. C. 2010. "Automated speech act classification in Arabic". *Subjetividad y Procesos Cognitivos*, 14: 284-292.
- (Shriberg et al., 2004) Shriberg E., R. Dhillon, S. Bhagat, J. Ang and H. Carvey. 2004. "The ICSI meeting recorder dialog act (MRDA) corpus". In *Proceedings of the 5th SIGDIAL Workshop on Discourse and Dialogue*.
- (Silvia et al., 2011) Silvia Q., V. Alexei and R. Giuseppe. 2011. "Simultaneous dialog act segmentation and classification from human-human spoken conversations". *ICASSP 2011*, 5596-5599.
- (Stolcke et al., 2000) Stolcke A., K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. 2000. "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech". In *Computational Linguistics 2000*. 26(3): 339-373.
- (Surendran et Levow, 2006) Surendran, D., & Levow, G. A. 2006. "Dialog act tagging with support vector machines and hidden Markov models. In *Ninth International Conference on Spoken Language Processing*.
- (Tavafi et al., 2013) Tavafi, M., Mehdad, Y., Joty, S., Carenini, G., & Ng, R. 2013. "Dialogue act recognition in synchronous and asynchronous conversations. In *Proceedings of the SIGDIAL 2013 Conference*, 117-121.
- (Tran et al., 2017) Tran, Q. H., Zukerman, I., & Haffari, G. 2017. "Preserving distributional information in dialogue act classification". In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2151-2156.
- (Trognon et Batt, 2007) Trognon, A. and Batt, M. 2007. "La Logique interlocutoire : un cadre théorique pour l'analyse psycho-sociale de l'usage du langage". *Cognition, santé et vie quotidienne*, 1, 183.
- (Verbree et al., 2006) Verbree, D., R. Rienks, and D. Heylen. 2006. "Dialogue Act Tagging using Smart Feature Selection; Results on Multiple Corpora". *Spoken Language Technology Workshop*, 2006. 70-73.
- (Wahlster, 2000) Wahlster, W. 2000. "Verbmobil: Foundations of Speech-To-Speech Translation". *Springer Science & Business Media*.
- (Walker et al., 2001) A Le, S. Lee, S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, G. Sanders, S. Sene, D. Stallard, S., and Whittaker. 2001. "DARPA Communicator

Dialog Travel Planning Systems: The June 2000 Data Collection". In *The European Conference on Speech (EuroSpeech'01)*.

(Webb et al., 2005) Webb N., M. Hepple and Y. Wilks. 2005. "Dialogue act classification based on intra-utterance features". In *Proceedings of the AAAI Workshop on Spoken Language Understanding*, Pittsburgh, PA.

(Webb et al., 2008) Webb, N., T. Liu, M. Hepple, and Y. Wilks. 2008. "Cross-Domain DialogueAct Tagging". In *proceedings of the 6th International Conference on Language Resources and Evaluation (LREC2008)*, Marrakech, Morocco.

## LISTE DE PUBLICATIONS

Articles dans des revues internationales		
<p>[1] <b>Samira Ben Dbabis</b>, Hatem Ghorbel and Lamia Hadrich Belguith, "Sequential dialogue act recognition for Arabic argumentative debates". <i>Procesamiento del Lenguaje Natural (SEPLN)</i>, 60, 53-60., 2018.</p> <p><a href="https://www.scimagojr.com/journalsearch.php?q=21100195304&amp;tip=sid&amp;clean=0">https://www.scimagojr.com/journalsearch.php?q=21100195304&amp;tip=sid&amp;clean=0</a></p>	<b>Indexation</b>	SJR
	<b>Impact Factor</b>	0.24
	<b>Classe</b>	Q2
Articles dans des conférences internationales		
<p>[2] <b>Samira Ben Dbabis</b>, Hatem Ghorbel, Lamia Hadrich Belguith and Mohamed Kallel, "Automatic dialogue act annotation within Arabic debates". <i>16<sup>th</sup> International Conference on Intelligent Text Processing and Computational Linguistics(CICLing)</i>, April 14-20, 2015, Cairo, Egypt.</p>	<b>Indexation</b>	Scopus / SJR/Core DBLP / Libra ISI Web of Science
	<b>Classe(Core)</b>	B
<p>[3] <b>Samira Ben Dbabis</b>, Boutheina Rguui, Hatem Ghorbel and Lamia Hadrich Belguith, "Utterance Segmentation Using Conditional Random Fields". <i>27<sup>th</sup> International Business on Information Management Association (IBIMA)</i>, May 3-5, 2016, Milano, Italy.</p>	<b>Indexation</b>	Scopus / SJR/Core ISI Web of Science Engineering Village
	<b>Classe(Core)</b>	B
Articles dans des "Workshops"		
<p>[4] <b>Samira Ben Dbabis</b>, Fatma Mallek, Hatem Ghorbel and Lamia Hadrich Belguith, " Dialogue Acts Annotation Scheme within Arabic discussions". <i>Sixteenth Workshop on the Semantics and Pragmatics of Dialogue(SemDial)</i>, September 19-21, 2012. Paris, France.</p>		

---

# ANALYSE DES DISCUSSIONS HOMME-HOMME EN ARABE

---

**Résumé :** Cette thèse s'intègre dans le cadre de l'analyse des conversations humaines et plus précisément aux débats politiques télévisés arabes. L'objectif principal est d'étudier les aspects argumentatifs présents dans ce type de discussions afin de dégager les conflits, l'échange d'opinions, les acceptations et les rejets des idées discutées tout au long du dialogue. Pour se faire, nous nous focalisons sur la reconnaissance des actes de dialogue qui projettent l'intention de l'interlocuteur et le contexte d'énonciation de son discours. Nous avons proposé un schéma d'annotation qui reflète la structure communicative et argumentative des débats politiques. La méthode proposée se base principalement sur les modèles CRF qui tiennent en compte des dépendances entre les mots d'un même énoncé. Ces modèles ont été appliqués pour les deux tâches de segmentation et d'annotation des actes de dialogue. Les résultats d'évaluation obtenus sont prometteurs et confirment l'efficacité des modèles CRF pour les tâches de segmentation et d'étiquetage des données séquentielles.

**Mots clés :** Reconnaissance des actes de dialogue, débats politiques arabes, schéma d'annotation, actes argumentatifs, modèles CRF.

**Abstract:** This thesis deals with human conversations analysis specifically the Arabic political debates. The main objective is to study the argumentative aspects present in such discussions in order to identify conflicts, the exchange of opinions, acceptances and rejections of the ideas discussed throughout the dialogue. To perform this task, we focus on the recognition of dialogue acts which project the intention of the speaker and the context of utterance. We proposed an annotation scheme that reflects the communicative and argumentative structure of political debates. The proposed method is based mainly on the CRF models that take into account dependencies between words in the same utterance. These models were applied to the two tasks of segmentation and annotation of dialogue acts. The evaluation results are promising and confirm the effectiveness of CRF models for tasks of segmentation and labeling of sequential data.

**Keywords:** Dialogue acts recognition, Arabic political debates, annotation scheme, argumentative acts, CRF models.