Reconnaissance des entités nommées à partir de Wikipédia arabe

Fatma Ben Mesmia

École Doctorale de Mathématiques, Informatique, Science et Technologie de la Matière

Faculté des Sciences Mathématiques, Physiques et Naturelles de Tunis

Université Tunis El Manar

Laboratoire de MIRACL (Multimedia, InfoRmation systems and Advanced Computing Laboratory)

Thèse

Reconnaissance des entités nommées à partir de Wikipédia arabe : Application à la découverte des relations sémantiques

Présentée pour obtenir le grade de

DOCTEUR EN INFORMATIQUE

Par

Fatma Ben Mesmia Chaabouni

Soutenue le 11/04/2019

Devant le jury composé de

Mohamed Mohsen Gammoudi	Professeur à l'Institut Supérieur des Arts et Multimédia Manouba	Président
Mounir Zrigui	Professeur à la Faculté des Sciences de Monastir	Rapporteur
Éric Laporte	Professeur à l'Université Paris-Est Marne-la-Vallée	Rapporteur
Afef Kacem	Maître de conférences à l'Ecole Nationale Supérieure d'ingénieurs de Tunis	Examinatrice
Kais Haddar	Professeur à la Faculté des Sciences de Sfax	Directeur de thèse
Denis Maurel	Professeur à l'Université de Tours	Co-directeur de thèse
Nathalie Friburger	Maître de conférences à l'Université de Tours	Invité

Publications

[2017] Ben Mesmia F., Zid F., Friburger N., Haddar K. and Maurel D. ASRextractor: A Tool extracting Semantic Relations between Arabic Named Entities. 3rd International Conference on Arabic Computational Linguistics, ACLing, Dubai, United Arab Emirates. Une partie du numéro special : Arabic Computational Linguistics. Procedia Computer Science. Vol 117. p. 55-62. H Index = 29. SJR = 0.27 en 2016.

Lien du papier : https://www.sciencedirect.com/science/article/pii/S1877050917321804

[2017] Ben Mesmia F., Friburger N., Haddar K. and Maurel D. CasANER: Arabic Named Entity Recognition Tool. Intelligent Natural Language Processing: Trends and Applications. Springer, Cham, 2018. p. 173-198. eBook ISBN : 978-3-319-67056-0. Studies in Computational Intelligence book series. H Index = 37. SJR = 2.5 en 2016.

Lien du papier : https://link.springer.com/chapter/10.1007/978-3-319-67056-0_10

[2017] Ben Mesmia F., Bouabidi K., Friburger N., Haddar K. and Maurel D. Extraction of Semantic Relation between Arabic Named Entities Using Different Kinds of Transducer Cascades. 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017), Budapest, Hungary. Springer LNCS - Lecture Notes in Computer Science. H Index = 251. SJR = 03 en 2016. Rang B.

Lien du papier accepté : https://www.cicling.org/2017/accepted.html

[2016] Ben Mesmia F., Friburger N., Haddar K. and Maurel D. Recognition and TEI annotation of Arabic Events Using Transducers. 2nd International Conference on Arabic Computational Linguistics (AcLing). Konya, Turquie.

Lien du papier accepté : https://www.cicling.org/2016/accepted.html

[2015] Ben Mesmia F., Friburger N., Haddar K. and Maurel D. Construction d'une cascade de transducteurs pour la reconnaissance des dates à partir d'un corpus Wikipédia. Colloque pour les Étudiants Chercheurs en Traitement Automatique du Langage naturel et ses applications, pages 8-11.

Lien du papier : https://web2.qatar.cmu.edu/~wajdiz/cectal2015/Actes_CEC-TAL2015.pdf

[2015] Ben Mesmia F., Friburger N., Haddar K. and Maurel D. Transducer cascade for an automatic recognition of Arabic Named Entities in order to establish links to free resources. First International Conference on Arabic Computational Linguistics (ACLing). pp 61-67.

Lien du papier : https://ieeexplore.ieee.org/document/7422281/

[2015] Ben Mesmia F., Friburger N., Haddar K. and Maurel D. 2015. Arabic Named Entity Recognition Process using Transducer Cascade and Arabic Wikipedia. Proceedings of Recent Advances in Natural Language Processing, pages 48–54, Hissar, Bulgaria. H Index = 15. SJR = 0.17 en 2016. Rang C.

Lien du papier : http://www.aclweb.org/anthology/R15-1007

Table des matières

Publications I

Table des matières III

Liste des figures XII

Liste des tableaux XV

Liste des abréviations XVI

Introduction générale 1

1. Entités nommées et relations sémantiques 2

2. Reconnaissance des entités nommées 2

3. Corpus et langue arabe 3

4. Contributions 3

5. Manuscrit 4

Partie 1 : Etat de l'art 7

Chapitre 1 : Extraction des entités nommées et des relations sémantiques 8

1. Définitions d'une entité nommée 10

1.1. Définition d'une entité nommée de Ehrmann 10

1.2. Définition d'une entité nommée de Poibeau 10

1.3. Définition d'une entité nommée du projet Quaero 10

1.4. Définition d'un évènement du projet Quaero 11

1.5. Définition d'un évènement médicale de Tourille 11

1.6. Discussion des définitions proposées 11

2. Catégorisation d'une entité nommée 12

2.1. Catégorisation des conférences MUC 12

2.2. Catégorisation des conférences 13

2.3. Catégorisation des campagnes d'évaluation 13

2.4. Discussions 14

3. Approches de REN 15

3.1. Approche symbolique 15
3.2. Approche statistique 15
3.3. Approche hybride 16
4. Systèmes de REN existants 16
4.1. Systèmes symboliques de REN 16
4.1.1. Système NERA 2.0 16
4.1.2. Système de Aboaoga et Aziz 17
4.1.3. Système de Fehri 17
4.1.4. Système RENAM 18
4.1.5. Système de Btoush et al 19
4.2. Systèmes statistiques de REN 19
4.2.1. Système de Darwish et Gao 19
4.2.2. Système de Kanya et Ravi 20
4.2.3. Système de Salleh et al 20
4.2.4. Système de Yao et al 21
4.2.5. Système de Mohammed et Omar 21
4.3. Systèmes hybrides de REN 21
4.3.1. Système de Zribi et al 22
4.3.2. Système de Küçük et Yazıcı 22
4.3.3. Système de Shaalan et Oudah 22
.4.3.4 Système de Hkiri et al 23
4.3.5. Système de Sharma et al 23
4.3.6. Système de Ramesh et Sanampudi 24
4.4. Discussions 24
5. Relations sémantiques entre les entités nommées 26
5.1. Apparition de la notion de RS pour les EN 27
5.2. Systèmes symboliques d'extraction de RS 27

5.2.1. Système de Ben abacha et al 27
5.2.2. Système IMAIOS 28
5.2.3. Système de Ghamnia 28
5.2.4. Système de Ezzat 28
5.2.5. Système de Ben Hamadou et al 29
5.3. Systèmes statistiques d'extraction de RS 29
5.3.1. Système de Abd El-Salam et al 29
5.3.2. Système de Tourille et al 30
5.4. Systèmes hybrides d'extraction de RS 30
5.4.1. Système de Lahbib et al 30
5.4.2. Système de Boujelben et al 31
5.5. Discussions 31
6. Wikipédia 32
6.1. Volume arabe de la Wikipédia 32
6.2. Travaux exploitant la Wikipédia 33
7. Norme d'annotation TEI 34
7.1. Aperçu sur la TEI 35
7.2. Travaux exploitant la TEI pour la langue arabe 35
8. Conclusion 36
Chapitre 2 : Aperçu sur les automates et les transducteurs 38
1. Automates 40
2. Automate à nombre fini d'états 40
2.1. Représentation formelle d'un automate à nombre fini d'états 40
2.2. Automate à nombre fini d'états déterministe ou non déterministe 41
2.3. Minimisation des automates à nombre fini d'états 42
2.4. Union des automates à états finis 43
2.5. Création d'un conjugueur à la base des automates à états finis 43

2.6. Création d'un traducteur à la base des automates à états finis 44
3. Réseau de transitions 44
3.1. Réseau de transition simple (RTS) 44
3.2. Réseau de transition récursif (RTR) 44
3.3. Réseau de transition augmenté (RTA) 46
4. Transducteurs à nombre fini d'états 46
4.1. Définition formelle d'un transducteur 46
4.2. Représentation graphique d'un transducteur 47
4.3. Traduction automatique à la base des transducteurs 47
4.4. Résolution de l'agglutination à la base des transducteurs 48
4.5. Reconnaissance des syntagmes dans des EN via des transducteurs 48
4.6. Segmentation des textes à la base des transducteurs 49
5. Cascade de transducteurs 50
6. Systèmes basés sur les cascades de transducteurs 50
6.1. Cascades de transducteurs pour l'Extraction d'Information 51
6.2. Cascades de transducteurs pour la fouille de texte 51
6.3. Cascade de transducteurs pour la segmentation de parole 52
7. Unitex 52
7.1. Transducteurs sous Unitex 52
7.1.1. Notion de variables dans les transducteurs 53
7.1.2. Notion de répétition dans les transducteurs 53
7.1.3. Notion de contexte négatif 54
7.1.4. Notion de filtre et mode morphologique 54
7.2. Création d'une cascade de transducteurs sous Unitex 55
7.2.1. CasSys 55
7.2.2. Application et modes de passage d'une cascade de transducteurs ... 56
7.3. Comparaison entre Unitex et NooJ 57

8. Conclusion 58

Partie 2 : Etude linguistique 60

Chapitre 3 : Typologie des entités nommées arabes 61

1. Identification des ENA 62

1.1. Définition d'ENA retenue 63

1.2. Hiérarchie d'ENA établie 63

2. Catégorisation d'ENA 65

2.1. Catégorie Date 65

2.2. Catégorie Nom de personne 68

2.2.1. Al-ism 68

2.2.2. Al-kuniyah 68

2.2.3. Al-nasab 68

2.2.4. Al-laqab 68

2.2.5. Al-nisba 68

2.2.6. Les formes de noms de personnes identifiées 69

2.3. Catégorie Nom de lieu 69

2.3.1. Nom de lieu absolu 69

2.3.2. Nom de lieu géographique 70

2.3.3. Nom de lieu relatif 71

2.4. Catégorie Organisation 72

2.5. Catégorie Evènement 73

2.5.1. Evènement politique 73

2.5.2. Evènement culturel 74

2.5.3. Evènement religieux 74

3. Imbrication des ENA 74

4. Conclusion 75

Chapitre 4 : Typologie des relations sémantiques 76

1. Définition des RS retenue 77
2. Identification des RS reliant des ENA catégorisées 78
2.1. Synonymie 78
2.2. Méronymie 79
2.3. Accessibilité 79
2.4. Fonctionnelle 80
2.5. Proximité 81
2.6. Appartenance 81
2.7. Date de naissance 82
2.8. Date de décès 82
2.9. Année de fondation 83
2.10. Familiale 84
2.11. Origine 85
2.12. Equivalence de dates 85
2.13. Date politique 86
2.14. Place politique 86
3. Recherche des RS reliant les ENA à des informations significatives 87
3.1. RS reliant des ENA à des autres ENA non catégorisée 87
3.2. RS reliant des ENA à des mots pertinents 87
3.2.1. Nationalité 87
3.2.2. Religion 88
3.2.3. Profession 88
4. Hiérarchie de types de RS établie 89
5. Conclusion 90
Partie 3 : Démarche proposée 91
Chapitre 5 : Démarche proposée pour la reconnaissance des entités nommées arabes 92
1. Démarche proposée pour reconnaître et annoter les ENA 93

2. Modélisation des dictionnaires 95
3. Grammaires proposées 96
3.1 Grammaire pour les mots déclencheurs 96
3.1.1 Mots déclencheurs pour les noms de personne 96
3.1.2 Mots déclencheurs pour les noms d'organisation 97
3.2 Grammaire proposée pour les ENA 98
4. Etablissement des transducteurs 99
4.1 Transducteurs traitant la langue arabe 100
4.1.1 Résolution de l'agglutination 100
4.1.2 Analyse syntagmatique 101
4.2 Transducteurs de reconnaissance et annotation des ENA 103
4.2.1 Transducteurs d'analyse 103
4.2.2 Transducteurs de filtrage 106
4.2.3 Transducteurs de généralisation d'étiquetage 107
4.3 Transducteurs de normalisation de l'annotation des ENA 109
5. Conclusion 111
Chapitre 6 : Démarche proposée pour l'extraction des relations sémantiques 113
1. Description de la démarche proposée pour extraire des RS 114
2. Etablissement des expressions régulières 116
2.1. Expression régulière associée à la RS Equivalence date 116
2.2. Expression régulière associée à la RS Date politique 117
2.3. Expression régulière associée à la RS Familiale 117
2.4. Expression régulière associée à la RS place politique 118
2.5. Exemples d'expressions régulières pour les types de RS identifiées ... 118
3. Modélisation d'un dictionnaire sémantique 120
4. Etablissement des transducteurs d'analyse 121
4.1. Création des sous-graphes 121

4.2. Exploitation des variables 122
4.3. Exploitation du mode morphologique 124
4.4. Traitement de l'éloignement dans les transducteurs d'analyse 124
5. Importance des RS dans l'enrichissement d'un dictionnaire d'ENA 126
6. Conclusion 128
Partie 4 : Implémentation, expérimentation et évaluation 129
Chapitre 7 : Implémentation des systèmes CasANER et ASRextractor 130
1. Implémentation du système CasANER 131
2. Phase de prétraitement 132
2.1. Segmentation du corpus 133
2.2. Suppression des liens internes 134
3. Création et enrichissement des dictionnaires manuellement 135
3.1. Enrichissement des dictionnaires des noms propres existants 136
3.2. Création des nouveaux dictionnaires 137
3.3. Création des nouveaux dictionnaires automatiquement 137
3.3.1. Implémentation d'un extracteur 137
3.3.2. Processus de tri des noms communs issus de l'ATB 139
3.3.3. Mise en correspondance des traits de l'ATB et Unitex 140
3.3.4. Création des graphes de transformation et filtrage de traits 141
3.3.5. Création d'une cascade de transformation de traits 142
3.3.6. Création d'un dictionnaire d'adjectifs 142
4. Implémentation proprement dite 143
5. Implémentation du système ASRextractor 145
5.1. Création d'un dictionnaire sémantique prioritaire 146
5.2. Implémentation du système ASRextractor 147
6. Implémentation de la cascade de normalisation des ENA 148
7. Post-traitement relié aux ENA dans les RS 149

8. Conclusion 151

Chapitre 8 : Expérimentation et évaluation des systèmes CasANER et ASRextractor 152

1. Phase d'expérimentation 153

1.1. Expérimentation du système CasANER 154

1.2. Expérimentation du système ASRextractor 154

1.3. Expérimentation de la cascade de normalisation 155

1.4. Expérimentation de la cascade de récupération de la forme brute des ENA 156

2. Evaluation de CasANER pour toutes les catégories 157

3. Evaluation de CasANER par catégorie 158

4. Evaluation de ASRextractor pour tous les types de RS 158

5. Evaluation d'ASRextractor par type de RS 160

6. Évaluation CasANER sur ANERcorp et comparaison avec ANERsys 161

6.1. Présentation du corpus ANERcorp 161

6.2. Principe de l'évaluation et la comparaison 163

6.2.1. Prétraitements effectués sur ANERcorp 163

6.2.2. Création des graphes d'adaptation d'annotation 164

6.2.3. Création de la cascade d'adaptation de l'annotation 166

6.3. Evaluation et comparaison 167

6.4. Problèmes rencontrés lors de l'application de CasANER sur ANERCorp 169

6.5. Enrichissement de ANERcorp grâce au système CasANER 170

7. Conclusion 170

Conclusion générale 172

Bibliographies 177

Liste des figures

Figure 1. Article de la Wikipédia arabe et sa structure interne 33
Figure 2. Diagramme de transition d'automate à nombre fini d'états 41
Figure 3. Principe d'une cascade des transducteurs 50
Figure 4. Ajout des dictionnaires pour le mode morphologique 54
Figure 5. Interface de l'outil CasSys 55
Figure 6. Application d'une cascade de transducteurs 56
Figure 7. Modes de passage d'une cascade de transducteurs 56
Figure 8. Hiérarchie d'ENA établie 63
Figure 9. Sous-hiérarchie décrivant les sous-catégories d'un nom de lieu absolu 64
Figure 10. Sous-hiérarchie décrivant les sous-catégories : relative et géographique ... 64
Figure 11. Sous-catégories associées à un nom d'organisation 72
Figure 12. Hiérarchie de types de RS établie 89
Figure 13. Etapes de la démarche proposée pour reconnaître et annoter les ENA 94
Figure 14. Transducteur résolvant l'agglutination dans le cas d'un nom commun 101
Figure 15. Transducteur traitant un état construit 101
Figure 16. Transducteur traitant un syntagme adjectival 102
Figure 17. Transducteur traitant un syntagme prépositionnel 103
Figure 18. Reconnaissance une date composée par une saison et une année 104
Figure 19. Transducteur reconnaissant un nom de personne 104
Figure 20. Transducteur reconnaissant un nom de ville 105
Figure 21. Transducteur reconnaissant un nom de mer 106
Figure 22. Transducteur de filtrage reconnaissant un nom de personne 107
Figure 23. Transducteur de généralisation d'étiquetage pour les noms de personnes 108
Figure 24. Graphe d'étiquetage générique pour la catégorie Date 108
Figure 25. Architecture illustrative de la normalisation de l'annotation d'ENA 110
Figure 26. Normalisation de l'annotation d'une ENA sans élément 110
Figure 27. Normalisation de l'annotation d'une ENA sans élément 111
Figure 28. Processus d'extraction des RS entre les ENA 115
Figure 29. Expression régulière de la RS équivalence date 116
Figure 30. Expression régulière de la RS date politique 117
Figure 31. Expression régulière de la RS familiale 117
Figure 32. Expression régulière pour la RS place politique 118

Figure 33. Transducteur principal reconnaissant la synonymie 122
Figure 34. Transducteur reconnaissant la synonymie entre deux noms de ville 123
Figure 35. Extraction de la méronymie entre un évènement et une ville 123
Figure 36. Extraction de l'accessibilité entre les deux noms de lieux 124
Figure 37. Extraction de la relation d'appartenance entre les noms des montagnes .. 125
Figure 38. Entrées/sorties du système CasANER 132
Figure 39. Graphe de segmentation d'un texte arabe intégré sous Unitex 133
Figure 40. Interface facilitant la segmentation 134
Figure 41. Graphe de suppression des liens internes 134
Figure 42. Extrait d'un article de la Wikipédia arabe contenant des liens 135
Figure 43. Résultat de la suppression des liens 135
Figure 44. Liste des dictionnaires intégrés sous la plateforme Unitex 136
Figure 45. Processus de la création automatique de dictionnaires 138
Figure 46. Extrait du code de l'extracteur de noms communs 139
Figure 47. Graphe de transformation trait de l'ATB vers Unitex 141
Figure 48. Graphe de filtrage de trait 141
Figure 49. Extrait de la liste des noms communs issu de l'ATB 142
Figure 50. Extrait de la liste issu du processus de création 142
Figure 51. Extrait de la liste des adjectifs générée à partir de l'ATB 143
Figure 52. Architecture du système CasANER 144
Figure 53. Cascade de transducteurs représentant le système CasANER 145
Figure 54. Entrée/Sortie du système ASRextractor 146
Figure 55. Chargement du dictionnaire sémantique prioritaire 147
Figure 56. Cascade de transducteurs représentant le système ASRextractor 148
Figure 57. Cascade de normalisation de l'annotation des ENA en TEI 148
Figure 58. Exemple de normalisation de l'annotation des ENA 149
Figure 59. Graphe de récupération de la forme brute d'une ENA sans type 150
Figure 60. Graphe de récupération de la forme brute d'une ENA avec type 150
Figure 61. Cascade de récupération de la forme brute d'ENA 151
Figure 62. Extrait d'un fichier de sortie de CasANER 154
Figure 63. Extrait d'un fichier de sortie de ASRextractor 155
Figure 64. Annotation normalisée d'une ENA 155
Figure 65. Résultat de récupération de la forme brute des ENA 156
Figure 66. Exemple d'erreur liée à la structure de la Wikipédia 160

Figure 67. Exemple d'erreur générée par le système CasANER 160
Figure 68. Principe de l'évaluation du système CasANER sur le ANERcorp 163
Figure 69. Prétraitements effectués sur ANERcorp 164
Figure 70. Graphe de transformation de l'annotation d'une ENA composée 165
Figure 71. Graphe d'une deuxième transformation d'annotation d'une ENA 165
Figure 72. Graphe d'élimination de l'annotation d'un mot quelconque 166
Figure 73. Cascade d'adaptation de l'annotation d'ANERcorp 166
Figure 74. Architecture de la REN effectuée par CasANER sur ANERcop brut 167
Figure 75. Graphe de transformation ajouté à la cascade de synthèse 168
Figure 76. Nouvelle forme de la cascade de normalisation..... 168

Liste des tableaux

Tableau 1. Récapitulation sur les définitions d'EN 11
Tableau 2. Récapitulation sur les catégorisations antérieures 14
Tableau 3. Récapitulation sur les travaux de REN antérieurs 25
Tableau 4. Récapitulation sur les travaux antérieurs d'extraction des RS 31
Tableau 5. Comparaison entre Unitex et NooJ 57
Tableau 6. Sous-catégories associées à un nom de lieu relatif 71
Tableau 7. Mise en correspondance des traits de dictionnaires pour la REN 95
Tableau 8. Expressions régulières créées pour quelques catégories d'ENA 99
Tableau 9. Exemples d'expression régulière pour les types de RS identifiées 119
Tableau 10. Traits associés au dictionnaire sémantique 120
Tableau 11. Couverture de dictionnaires avant et après l'enrichissement 136
Tableau 12. Couverture de dictionnaires créés manuellement 137
Tableau 13. Extrait de processus de mappage 140
Tableau 14. Couverture des dictionnaires créés automatiquement 143
Tableau 15. Evaluation de CasANER avec les métriques de performance 157
Tableau 16. Evaluation de CasANER par catégorie 158
Tableau 17. Evaluation de ASRextractor avec les métriques de performance 159
Tableau 18. Evaluation d'ASRextractor par type de RS 161
Tableau 19. Balises d'annotation utilisées par ANERsys 162
Tableau 20. Comparaison entre ANERsys et CasANER 169

Liste des abréviations

AC : Arabe Classique

ACE : Automatic Content Extraction

AD : Arabe Dialectal

ASM : Arabe Standard Moderne

CoNLL : Conference On Natural Language Learning

CR : Container Relation

DCT : Document Creation Time

DR : Document creation time Relation

EI : Extraction d'Information

EL : Entity Linking

EN : Entité Nommée

ENA : Entité Nommée Arabe

ESTER : Evaluation Systèmes de Transcription d'Emission Radiophoniques

EVALITA : Evaluation of NLP and Speech Tool for Italian

GATE : General Architect for Text Engeneering

IREX : Information Retrieval and Extraction Exercice

ISO : International Organization for Standarization

LMF : Lexical Markup Framework

MERLOT : Multimedia Educational Resource for learning and Online Teaching

MET : Multilingual Entity Task

MUC : Message Understanding Conference

RDF : Resource Description Framework

REN : Reconnaissance des Entités Nommées

RENAM : Reconnaissance des Entités Nommées Amazighes

RS : Relation Sémantique

SGML : Standard Generalized Markup

SVM : Super Vector Machine

TAC : Text Analysis Conference

TAL : Traitement Automatique des Langues

TALN : Traitement Automatique des langues Naturel

TEI : Text Encoding Initiative

THYME : Temporal History of Your Medical Events

UMLS : Unifed Medical Language System

XML : eXtensible Markup Language

Introduction générale

1. Entités nommées et relations sémantiques

Une entité nommée (EN) est une unité linguistique qui peut se composer par des éléments lexicaux ayant une signification interprétative dans un document textuel. Divers chercheurs ont tenté à valoriser et attribuer des définitions précises aux EN comme [Erhmann 2008] pour qui « Étant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus ». Ces chercheurs ont fait des études sur la base d'une variété de critères (référence, unicité, etc.). Cependant, les définitions d'EN établies ont évolué au cours du temps selon le domaine d'étude (spécifique ou générique). Les EN peuvent être ambigüe. Elles sont reliées sémantiquement entre elles. La découverte de ces liens enrichit les documents textuels qui les contiennent. Les relations sémantiques (RS) entre les EN améliorent la désambigüisation, et, donc, la recherche d'information (RI).

2. Reconnaissance des entités nommées

La reconnaissance des entités nommées (REN) reste encore une piste de recherche intéressante malgré qu'elle ait débuté depuis les années 1980. L'évolution de la REN se nourrit de nouvelles circonstances dûs à l'explosion du Web et aux grandes masses de données disponibles ayant besoin d'être analysées. La REN associe l'analyse syntaxique et l'analyse sémantique. L'analyse syntaxique permet une délimitation des EN (longueur et limite). L'analyse sémantique englobe principalement l'annotation des EN en des catégories prédéfinies, en ajustant éventuellement le niveau de granularité. La REN participe en grande partie à enrichir sémantiquement les documents textuels. Cet enrichissement peut augmenter la pertinence des résultats retournés par les systèmes questions-réponses ou encore par les moteurs de recherche.

Par ailleurs, la REN est un préliminaire à la recherche des relations sémantiques (RS) reliant les ENA. Cette extraction des RS se base à nouveau sur une analyse syntaxique et sur une analyse sémantique. L'analyse lexicale permet d'étudier le contexte d'apparition des RS. L'analyse sémantique permet d'annoter les RS détectées. L'extraction des RS permet un retour sur le processus de REN déjà effectué en révélant soit des informations manquantes, soit des erreurs. Elle permet également de revoir et compléter la hiérarchie des catégories d'EN.

La séparation du processus de REN et de celui d'extraction des RS augmente le temps d'exécution. Cependant, le couplage de ces deux processus va augmenter le nombre d'erreurs. Il est donc préférable de les séparer.

3. Corpus et langue arabe

Les ressources textuelles ayant des informations normalisées sont les plus interrogées par les moteurs de recherche sur le Web, car les mieux accessibles. Pour cette raison, la normalisation de l'annotation des EN et des RS permet d'obtenir des ressources linguistiques structurées extensibles et interopérables. L'utilisation d'un standard d'annotation assure ainsi la pérennité de l'exploitation de ces ressources.

Nous utiliserons comme corpus le volume arabe de l'encyclopédie Wikipédia qui est une ressource libre. La Wikipédia possède un volume arabe très riche en termes d'EN arabes (ENA) qui possèdent une fréquence d'apparition très importante et une présence dans divers contextes.

Les spécificités de la langue arabe posent divers problèmes dus à des phénomènes linguistiques complexes. L'agglutination, qui fait partie de l'analyse morphologique, est parfois présente dans le contexte d'apparition des ENA et des RS ou bien en leur sein. La non voyellation complique l'analyse lexicale car le texte traité est fortement ambigu. L'absence de capitalisation rend la détection des ENA plus compliquée. De plus, les travaux d'annotation existant pour l'arabe utilisent des balises définies localement. Généralement, ces corpus annotés obtenus ne peuvent pas être évalués par des outils associés à des standards d'annotation (tels que XML ou TEI).

4. Contributions

Dans ce contexte, notre objectif principal consiste à effectuer une reconnaissance des ENA à partir d'un corpus extrait de la Wikipédia arabe. Nous voulons ensuite exploiter le corpus issu de la REN pour la découverte des RS qui relient les ENA. Plusieurs chercheurs s'intéressent à réaliser la REN et l'extraction des RS pour la langue arabe.

Voici nos spécificités qui rendent originale notre travail sur la langue arabe :

1. Tout d'abord, pour l'annotation des ENA et les RS, nous choisissions la norme TEI pour avoir des résultats normalisés. La représentation standardisée des ENA et des RS, basée sur la TEI, facilite l'extensibilité et l'interopérabilité des ressources obtenues.
2. Puis nous exploitons le formalisme des machines à nombre fini d'états, qui utilisent des technologies avancées, comme les filtres morphologiques.
    a. En premier lieu, nous construisons un système, appelé CasANER, dédié à la reconnaissance des ENA. Pour ce faire, nous commençons par l'analyse de corpus extraits de la Wikipédia arabe afin d'attribuer à toutes les ENA apparues les catégories adéquates. Cette analyse nous permet de créer un ensemble de règles d'extraction via des indicateurs déclenchant la reconnaissance des ENA. Les règles créées vont être spécifiées à l'aide de trois ensembles de transducteurs

(analyse, filtrage et généralisation d'étiquetage) qui seront ordonnés dans une cascade selon un ordre de passage établi. Au système CasANER, nous associons un module de normalisation, à nouveau sous la forme d'une cascade de transducteurs pour finaliser l'annotation des ENA selon la norme TEI.

b. En deuxième lieu, nous proposons un deuxième système, intitulé ASRextractor, permettant d'extraire et annoter les RS entre les ENA à travers la version annotée du corpus de la Wikipédia arabe. Pour ce faire, nous analysons le nouveau corpus annoté en TEI par le système CasANER afin de dégager toutes les RS possibles entre les ENA et de les classifier en plusieurs types. Toutes ces règles vont être spécifiées à l'aide d'une cascade de transducteurs passés dans un ordre défini. Comme précédemment indiqué, le système ASRextractor est associé à un module de normalisation basé sur une nouvelle cascade de transducteurs.

5. Manuscrit

Le présent manuscrit s'articule autour de quatre parties :

La première partie intitulée « État de l'art » est composée de deux chapitres :

Le premier chapitre étant « Extraction des entités nommées et de relations sémantiques ». Dans ce chapitre, nous allons commencer par la présentation de la notion d'EN à travers les différentes définitions proposées dans la littérature. Puis, nous allons décrire les catégorisations effectuées depuis les conférences MUC ainsi que les conférences et les campagnes d'évaluation qui se basent sur leur principe. Ensuite, nous allons aborder trois approches d'Extraction d'Information (EI) principales (symbolique, statistique et hybride). Pour chaque approche, nous allons présenter les systèmes de REN et d'extraction de RS associés. Après, nous allons présenter en détail la ressource libre Wikipédia et les travaux qui l'exploitent. Enfin, nous clôturons ce chapitre par une partie dédiée à la représentation de norme TEI et à son importance.

Le deuxième chapitre s'intitule « Aperçu sur les automates et les transducteurs ». Dans ce chapitre, nous allons rappeler la définition et le principe d'un automate ainsi celui d'automate à états finis avec une illustration de quelques travaux qui les exploitent. Vu qu'un automate peut déclencher plusieurs types d'opérations, ces dernières seront expliquées avec des exemples décrivant leur principe. Ensuite, nous allons expliquer le formalisme des transducteurs et de ses représentations graphiques dépendant et indépendant d'une plateforme et les réseaux de transitions. Puis, nous allons présenter le principe d'une cascade de transducteurs et les travaux associés. Finalement, nous allons citer quelques plateformes linguistiques qui sont dédiées au TAL et qui se basent sur la manipulation des transducteurs et l'organisation de leur ordre de passage à travers divers modes.

La deuxième partie est appelée « étude linguistique » et est composée de deux chapitres :

Le troisième chapitre s'intitule « Typologie des entités nommées arabes ». Dans ce chapitre, nous allons décrire notre phase de recherche des ENA à effectuer en nous basant sur un corpus d'étude extrait à partir de la Wikipédia arabe. Dans cette phase de recherche, nous allons donner la définition d'ENA retenue, ainsi que la catégorisation effectuée et, pour finir, les catégories et les sous-catégories identifiées.

Le quatrième chapitre, « Typologie des relations sémantiques », est dédié à la recherche des RS entre les ENA apparaissant dans notre corpus d'étude et à la représentation de la typologie des types de RS.

La troisième partie est appelée « Démarche proposée » composée de deux chapitres :

Le cinquième chapitre intitulé « Démarche proposée pour la reconnaissance des entités nommées arabes » est consacré à la description de la démarche que nous préconisons pour reconnaître et annoter les ENA. Cette démarche repose principalement sur divers ensembles de transducteurs traduisant les expressions régulières établies. Les transducteurs conçus possèdent des rôles variés à savoir l'analyse, l'annotation ou la normalisation. Pour cela, nous allons exploiter des fonctionnalités avancées à savoir l'utilisation des variables, de contexte négatif et du mode morphologique.

Le sixième chapitre intitulé « Démarche proposée pour l'extraction des relations sémantiques » vise à décrire l'extraction et l'annotation des RS entre les ENA. Cette démarche repose sur la création des ressources nécessaires (un ensemble d'expressions régulières et un dictionnaire sémantique prioritaire). Les expressions régulières seront transformées en transducteurs d'analyse. Nous utiliserons la notion des variables pour l'organisation de l'annotation de sortie et celle de contexte négatif pour éviter certains phénomènes linguistiques.

La quatrième partie appelée « Implémentation et évaluation » est composée de deux chapitres :

Le septième chapitre appelé « Implémentation et expérimentation » introduit l'implémentation de nos systèmes proposés CasANER et ASRextractor grâce à la plateforme linguistique Unitex pour tirer profit de ses fonctionnalités avancées. Etant donné que les systèmes proposés se basent sur des cascades de transducteurs, nous allons exploiter l'outil CasSys intégré à Unitex. Les textes exploités vont subir des prétraitements accomplis via Unitex, dont la consultation de dictionnaires dont nous allons développer la couverture. Dans ce chapitre, nous abordons la description du module de normalisation pour transformer l'annotation des ENA reconnues par le système CasANER en une annotation conforme à la

norme TEI. Nous allons proposer également un post-traitement associé aux ENA figurant dans les RS pour enrichir les dictionnaires d'ENA.

Le dernier chapitre appelé « Expérimentation et évaluation des systèmes CasANER et ASRextractor » décrit l'évaluation des deux systèmes CasANER et ASRextractor après leur expérimentation. L'évaluation sera réalisée à travers les mesures de performance. Nous allons étudier les lacunes de chaque système proposé et localiser les points à améliorer ainsi que les erreurs rencontrées.

Enfin, nous clôturons cette thèse par une conclusion générale dans laquelle nous récapitulons le travail réalisé et nous présentons nos perspectives.

Partie 1 : Etat de l’art

Chapitre 1 : Extraction des entités nommées et des relations sémantiques

L'attribution d'une définition pour l'EN reste encore un intérêt de recherche car elle peut se baser divers critères d'identification d'une EN en cherchant leur portée sémantique. Ces critères diffèrent selon les besoins de l'application à réaliser et le domaine d'étude qui est devenu diversifié. Une définition d'EN peut parfois compliquer sa délimitation puisqu'il faut s'interroger sur quel est l'EN porteuse de sens à prendre en compte. La recherche d'une définition d'EN peut s'annoncer aussi via un processus de catégorisation qui constitue la décision d'assigner une catégorie adéquate à une EN et voir ce qu'elle recouvre en termes de sous-catégories. La typologie d'EN établie dans la littérature n'est pas toujours détaillée et elle englobe un nombre réduit de catégories ce qui reflète un manque de compromis de classification défini par les chercheurs pour toutes les langues ou pour une langue bien déterminée. L'absence d'une typologie développée d'EN a influé sur les résultats obtenus par les systèmes de REN ou d'extraction des RS existants malgré la diversité des approches fondamentales d'Extraction d'Information (EI) et la puissance des formalismes et des techniques qu'elle offre. Ces systèmes manquent aussi d'une mise à jour des documents textuels exploités car ils utilisent généralement des anciens corpus existants qui n'ont pas eu des extensions. L'annotation des EN et des RS est un processus non pris en considération par certains systèmes qui font recours à des préférences personnelles d'annotation ou aux balises associées à la conférence CoNLL. L'absence de la normalisation de l'annotation des EN et des RS empêchent l'obtention des informations sémantiques unifiées, compréhensibles et structurées. Il décrémente aussi la chance d'avoir des documents textuels interopérables et capables d'être intégrés dans le Web sémantique.

Le chapitre présent se divise en sept parties décrivant l'état de l'art que nous effectuons. Dans la première partie, nous présentons quelques définitions attribuées à l'EN et nous étudions la différence entre eux. La partie deux décrit les catégorisations d'EN depuis les conférences MUC et proposées soit par des conférences soit par des campagnes d'évaluation. Cette partie aborde aussi une discussion faite pour distinguer les points de différence entre les catégorisations d'EN que nous citons. Dans les trois parties qui suivent, nous présentons les approches fondamentales d'EI (symbolique, statistique et hybride). Pour chaque approche, nous présentons des systèmes de REN et d'extraction des RS entre les EN associés. Nous ajoutons également, une partie de discussion pour illustrer les points forts et les lacunes des systèmes étudiés. Dans la sixième partie, nous expliquons la ressource libre Wikipédia plus précisément son volume arabe et nous citons quelques travaux de recherche qui l'utilisent. Dans la dernière partie, nous discutons la norme d'annotation TEI en donnant un aperçu sur son principe et en citant les travaux qui l'exploitent en langue arabe.

1. Définitions d'une entité nommée

Dans le domaine d'EI, plusieurs chercheurs ont tenté à valoriser et attribuer une définition d'une EN. La définition attribuée peut faciliter l'identification de ces EN dans des corpus pour déterminer leur limite et leur aspect sémantique. L'analyse des EN offre l'opportunité de faciliter la compréhension des documents textuels. Cependant, les définitions proposées différent selon des critères tels que la référence, l'unicité et le domaine d'étude. Dans la section courante, nous citons des définitions des EN proposées dans la littérature.

1.1. Définition d'une entité nommée de Ehrmann

Dans le but de savoir la capacité d'une EN à renvoyer un référent unique, Ehrmann dans [Ehrmann, 2008] a défini une EN comme suit : « *On appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le* corpus ». Selon cette définition, l'EN est considérée ainsi comme une forme d'expression qui recouvre une réalité caractérisable par rapport à la référence. Une EN peut être alors un nom propre comme « Victor Hugo » ou une expression temporelle (heure, journée, année, etc.). Cependant, les expressions ayant une description incomplète ne peuvent pas être des EN comme « le président de la république ». Cela revient à poser la question quel pays préside-t-il ? alors que « le président de la république française » peut-être une EN.

1.2. Définition d'une entité nommée de Poibeau

Avec plus de précision par rapport aux définitions précédentes, Poibeau dans [Poibeau, 2011] a considéré une EN comme étant « *Les types d'unités lexicales particuliers qui font référence à une entité du monde concret dans certains domaines spécifiques notamment humains, sociaux, politiques, économiques ou géographiques et qui ont un nom (typiquement un nom propre ou un acronyme)* ». L'EN définie couvre non seulement les noms propres (Gustave Flaubert), mais aussi des EN plus complexes comme les expressions ; de temps (18 mars 1989, date de naissance), de quantité (Cinq cents mètre, gradeur physiques).

1.3. Définition d'une entité nommée du projet Quaero

Le modèle d'annotation Quaero dans [Grouin et al., 2011] a proposé une nouvelle définition étendue définie comme suit : « *Une expression qui peut ne pas contenir un nom propre et elle peut être structurée via des sous-catégories et des composantes* ». Selon cette définition, une EN peut être assigner à une catégorie comme *nom de personne* comme « Bertrand Delanoë ». Une catégorie peut avoir des sous-catégories comme la catégorie *fonction* ayant la sous-catégorie *métier* comme « maire de Paris » (le ministère des affaires étrangères ayant la catégorie *organisation* et la sous-catégorie *administration*).

1.4. Définition d'un évènement du projet Quaero

Le modèle d'annotation Quaero admettait également une nouvelle définition pour une EN exprimant un événement. Selon [Arnulphy et Tanier, 2013], la définition est la suivante : « U*n évènement est ce qui se produit, doit arriver ou ne s'est pas réalisé. L'événement est ancré dans la temporalité* ». L'EN « la coupe du monde » peut être retenue par la définition proposée. Les expressions de durée peuvent aussi désigner un évènement comme « ces trois heures de musique écoutées comme en rêve ».

1.5. Définition d'un évènement médicale de Tourille

La nature des corpus peut faire varier les définitions d'une EN mais la qualité des corpus représentant le même domaine peut engendrer une légère différence entre ces définitions. Dans cette optique, [Tourille et al., 2017] ne se basaient pas sur les définitions précédentes attribuées à une EN médicale mais ils ont proposé une nouvelle définition liée juste à un évènement. Pour cette raison, les auteurs ont défini une EN médicale exprimant un évènement comme suit : « *Un événement médical est tout ce qui pourrait intéresser le calendrier clinique d'un patient* ». Selon la définition, un évènement ne pourrait être qu'une procédure médicale d'une maladie ou d'un diagnostic.

1.6. Discussion des définitions proposées

Les définitions d'EN précédentes étaient proposées après des études analytiques sur des corpus. Les EN étaient considérées comme des unités linguistiques caractérisées par des significations pertinentes. Dans le tableau 1, nous présentons les points forts et faibles de chaque définition déjà citée.

Tableau 1. Récapitulation sur les définitions d'EN

Définitions	Points forts	Points faibles
[Erhmann, 2008]	Entité ayant une référence unique et autonome	Pas d'indication sur la nature du domaine d'étude
[Poibeau, 2011]	Unité lexicale particulière et concrète L'EN peut être une expression complexe	Dépendance de domaine particulier
[Grouin et al., 2011]	Extension de l'EN pour être une expression sans nom propre Structuration en sous-catégories et composantes	Problème d'incohérence d'annotation Confusion entre certaines composantes
[Arnulphy et Tanier, 2013]	Définition claire et exhaustive pour l'EN évènement	Exclusion de certaines expressions d'évènement (les phénomènes climatiques)
[Tourille et al., 2017]	Définition approfondie pour un évènement médicale	Catégorisation restreinte Dépendance de domaine

D'après le tableau illustré ci-dessus, les EN être classiques (un nom de personne, un nom de lieu, etc.) ou spécifiques. Les EN spécifiques dépendent d'un domaine d'étude particulier comme les EN médicales. Toutes les définitions citées possèdent le même objectif qui est la délimitation d'une EN. Cependant, Les chercheurs ont commencé par des définitions d'une EN générique indépendamment de sa classe d'appartenance pour aller à des définitions propres à une classe définie d'avance (évènement). Cette spécification apparaît progressivement avec l'utilisation des corpus dépendant d'un domaine particulier (médical). La définition du projet Quaero prend en compte les expressions qui ne contiennent pas des noms propres alors que celle de Erhmann ne la considère que lors de la présence de son référent dans le même corpus.

Les définitions déjà citées ont été exprimées directement mais il existe d'autres définitions qui se représentent à travers une catégorisation d'EN. Dans la section suivante, nous présentons les catégorisations faites sur les EN depuis les conférences MUC.

2. Catégorisation d'une entité nommée

La catégorisation d'une EN est un processus visant à lui fournir une représentation adéquate pour élaborer une typologie d'EN utilisable par plusieurs applications de TAL. La catégorisation devient de plus en plus développée et détaillée selon la nature et la richesse de corpus étudiée. Cette catégorisation peut être aussi dépendante ou indépendante de domaine. La couverture d'un processus de catégorisation peut s'étendre et se raffiner jusqu'à qu'elle soit adoptée par d'autres langues.

Etant donné que la catégorisation a été proposée initialement par les conférences MUC, nous avons divisé cette section en trois parties dont la première est consacrée à décrire le principe de conférences déjà citées. De plus, nous présentons dans la deuxième partie les catégorisations faites par des conférences qui se basaient sur le principe de MUC tandis que celles propres aux campagnes d'évaluation qui adoptaient le même principe seront l'objectif de la troisième partie.

2.1. Catégorisation des conférences MUC

La catégorisation de l'EN a été proposée pour la première fois, pour la langue anglaise, dans la conférence MUC-6 [Grishman et Sundheim, 1996]. L'EN a été définie à travers 3 catégories principales qui sont *ENAMEX*, *TIMEX* et *NUMEX*. L'*ENAMEX* décrit les noms propres or le *TIMEX* est dédié aux expressions temporelles et le *NUMEX* représente les expressions numériques comme la monnaie et le pourcentage. La catégorie *ENAMEX* a été étendue pour classifier les noms propres via des sous-catégories : une sous-catégorie *personne* (nom de personne ou de famille), une sous-catégorie *location* (location politique ou géographique) et une sous-catégorie appelée *organisation* (nom d'entreprise). L'annotation de l'EN ayant ces

catégories a été faite via les balises SGML (Standard Generalized Markup Language). Les mêmes catégories de MUC-6 ont été prises en MUC-7 [Chinchor, 1997]. Une nouvelle catégorie appelée *airplane* a été créée dépendante de la nature du corpus (airplane crashes). Plusieurs conférences et projets qui adoptaient la catégorisation de MUC-6 ont raffiné les catégories proposées en se basant sur différents corpus.

2.2. Catégorisation des conférences

Il existe plusieurs conférences qui ont étendu la REN pour traiter de nouvelles langues autres que l'anglais. Parmi ces conférences, nous citons les deux conférences MET (Multilingual Entity Task) [Merchant et al., 1996] qui s'organisaient parallèlement aux MUC-6/7. Le MET a fourni une nouvelle opportunité évaluant le progrès de la REN en espagnol, en japonais et en chinois. Par conséquent, l'EN reconnue peut-être assignée aux catégories suivantes : *expressions* (nom de personne, organisation et location), *temps* (temps et date) et *expressions numériques* (pourcentage et monnaie). L'annotation a été faite via la norme SGML.

Durant le projet japonais appelé IREX (Information Retreivel and Extraction) [Grishman et Sundhein, 1996], une nouvelle catégorie *artifact* a été ajoutée (produits manufacturés, œuvres, prix, etc.). Dans ce projet, les auteurs montraient que la nouvelle catégorie était très difficile à détecter et elle avait diverses sous-catégories. Dans [Paik et al., 1996], les auteurs proposaient une nouvelle catégorisation pour augmenter le niveau de granularité et raffiner les catégories de MUC. L'extension de l'ancienne catégorisation a donné neuf classes principales avec 30 catégories au total. Les neuf classes principales sont *Geographic*, *Affiliation*, *Organization*, *Human*, *Document*, *Equipement*, *Scientific*, *Temporal* et *Miscellouneous*.

Dans CoNLL (Conference on Natural Language Learning) organisée en 2002 et 2003, les auteurs ont proposé quelques changements par rapport à la catégorisation de MUC [Tjong Kim Sang et De Meulder, 2003]. En fait, seulement les catégories *personne*, *location* et o*rganisation* ont été considérées et une catégorie appelée o*ptionnelle* était ajoutée. Ces catégories s'annotent en rajoutant l'affixe B pour désigner début (Begin) et I pour la suite (Inside) de l'EN reconnue.

2.3. Catégorisation des campagnes d'évaluation

Plusieurs campagnes d'évaluation adoptaient le même principe de MUC comme les campagnes ACE (*Automatic Content Extraction*) [Doddington et al., 2004] qui s'organisaient entre 1999 et 2008 et la campagne Ester [Gravier et al., 2004]. Ces campagnes d'évaluation visaient non seulement à rajouter de nouvelles catégories pour représenter une EN mais de raffiner également les catégories existantes. Ils tentaient à réaliser d'autres objectifs tels que

l'amélioration de l'évaluation, la production des corpus accessibles et à ouvrir de nouvelles perspectives à la tâche de REN.

Une nouvelle catégorisation différente de celles classiques a été proposée par le modèle d'annotation Quaero [Grouin et al., 2011]. L'EN a été étendue pour qu'elle puisse être structurée via des sous-catégories et des composantes. De même, la campagne d'évaluation EVALITA[1] (Evaluation of NLP and Speech Tools for Italian) organisée en 2007 est en évolution continue jusqu'à 2018 pour améliorer le développement et la diffusion des ressources textuelles pour l'italien en l'analysant via divers traitements [Basile et al., 2016].

2.4. Discussions

Les catégorisations que nous avons décrites diffèrent principalement selon la nature du corpus utilisé pour dégager les catégories d'EN. De plus, elles se distinguent selon la manière d'annoter choisie pour décrire les éléments d'EN. Dans le tableau suivant, nous rappelons les différentes catégorisations selon les deux critères déjà mentionnés.

Tableau 2. Récapitulation sur les catégorisations antérieures

Conférences/ Campagnes d'évaluation	Nombre de catégories	Choix d'annotation
MUC-6	3	SGML
MUC-7	4	SGML
MET	7	SGML
IREX	4	Non défini
Paik et al	30	Non défini
CoNLL	4	Balisage BIO
ACE	7	XML
Ester	6	Non défini
Quaero	11	XML
EVALITA	7	Non défini

Le tableau 2 montre la variété de catégories obtenues par les conférences et les campagnes d'évaluation. Les valeurs dans la deuxième colonne (nombre de catégories) n'évoluent pas avec le temps mais elles évoluent selon la richesse du corpus. Pour le choix d'annotation, certains travaux n'admettent pas une norme d'annotation et ils n'ont pas spécifié des spécifications à la manière d'annotation des EN. Cependant, d'autres travaux ont choisi d'annoter les catégories d'EN dégagées selon les normes SGML et XML.

La définition et la catégorisation des EN sont centrales à la tâche de REN. Cette dernière s'articule autour de trois parties principales dont la première est la catégorisation des EN. La

[1] http://www.evalita.it/2007

deuxième partie inclut le développement d'un système de REN qui peut adopter trois approches fondamentales (symbolique, statistique et hybride). La troisièmement partie est consacrée au choix de la norme d'annotation pour représenter et détailler l'EN et ses composantes. Ce choix doit assurer une annotation répondant aux besoins d'un système de REN. Dans la section suivante, nous commençons par la présentation des approches de REN dans la littérature.

3. Approches de REN

Il existe trois approches fondamentales de REN qui peuvent être fondées sur des démarches linguistiques ou non et qui sont nommées symbolique, statistique et hybride. Les trois approches permettent de réaliser les mêmes objectifs définis par une application tandis qu'elles admettent des principes différents. D'autres facteurs peuvent distinguer aussi ces approches comme l'acquisition de données et leur manipulation. La nature des informations étudiées ne fait pas la différence entre les approches déjà mentionnées. La section courante se divise en trois sous-sections dont chacune est dédiée à présenter une approche de REN.

3.1. Approche symbolique

L'approche symbolique repose sur la construction manuelle des règles à formaliser via des grammaires et à appliquer sur le corpus étudié. Les règles construites ont la forme des patrons dont ils reposent sur les caractéristiques ; morphologique, syntaxique et sémantique. Les règles peuvent être alimentées à l'aide des lexiques spécifiques. Etant donné que ces lexiques sont parfois non exhaustifs et ouverts alors la création des règles peut se baser sur des indices contextuels (des preuves externes déclenchant les EN) [Shaalan, 2010]. Dans le cas de la reconnaissance d'un nom de personne, une règle linguistique peut se composer d'un mot déclencheur (السيد/ Monsieur) plus deux mots se trouvant respectivement dans le lexique des prénoms et celui des noms de famille. L'approche symbolique repose sur des règles lisibles ce qui permet de cerner les erreurs rencontrées lors de leur application sur un texte. Les règles conçus assurent une précision car elles traitent la majorité des informations à chercher.

3.2. Approche statistique

L'approche statistique repose sur diverses techniques d'apprentissage qui se diffèrent au niveau du degré de supervision exigé. La supervision concerne l'intervention humaine pour l'étiquetage de l'ensemble de données dont l'objectif est de guider un modèle d'apprentissage déjà conçu. Au sein de l'approche statistique, nous distinguons trois types d'apprentissage ; supervisé, semi-supervisé et non supervisé. L'apprentissage supervisé consiste à créer un modèle à la base d'un ensemble de données annotées (nom de catégories) afin de classifier des nouvelles données. L'apprentissage semi-supervisé utilise deux ensembles de données ;

annotées et non annotées visant à améliorer la qualité d'apprentissage. Le dernier type est l'apprentissage non supervisé visant à créer des classes de données à partir d'un ensemble de données non annotées. Chaque type déjà cité fait recours à des algorithmes d'apprentissage qui seront appliqués pour entraîner le système élaboré [Nadeau et Sekine, 2007 ; Abdelrahman et al., 2010 ; Belainine, 2017].

3.3. Approche hybride

L'approche hybride est la combinaison des approches symbolique et statistique. La direction du flux de traitement peut être du symbolique vers le statistique ou vice versa. Autrement dit, les règles sont, soit écrites manuellement puis corrigées et améliorées automatiquement, soit apprises automatiquement puis révisées manuellement [Oudah et Shaalan 2012 ; Abuleil, 2006]. La combinaison de ces deux approches augmente la puissance descriptive des règles linguistiques d'une part et remède aux faiblesses d'apprentissage d'autre part. Cette approche aide à atteindre des améliorations importantes pour la performance des systèmes de REN.

4. Systèmes de REN existants

Les systèmes de REN s'appuient sur les trois approches déjà cités. Pour améliorer leur performance, ils profitent de la diversité des formalismes et techniques offerte par chaque approche. Dans ce qui suit, nous citons quelques systèmes de REN traitant plusieurs langues.

4.1. Systèmes symboliques de REN

Les systèmes symboliques de REN peuvent reconnaitre les EN selon deux formes de représentation des règles linguistiques possibles, soit sous selon le formalisme expressions régulières, soit sous celui de transducteurs à états finis [Mesfar, 2007]. Dans ce qui suit, nous citons quelques systèmes se basant sur les deux formalismes mentionnés.

4.1.1. Système NERA 2.0

Dans [Oudah et Shaalan, 2017], les auteurs ont proposé un système appelé « NERA 2.0 ». Ce système a réussi à augmenter la couverture des EN reconnues par un système appelé « NERA » élaboré par [Shaalan et Raza, 2009]. Le système NERA 2.0 conserve les mêmes catégories que l'ancien système qui sont *nom de personne, nom de lieu, organisation, date, heure, ISBN, le prix, mesure, numéros de téléphone* et *les noms de fichiers.* Sa mise en œuvre est effectuée dans le cadre FAST ESP (Continuité sur un système NERA) et elle comporte les mêmes composantes qu'un système appelé PERA. Ce dernier est proposé par [Shaalan et Raza, 2007] agissant sur les textes arabes. Cependant, le système PERA était dédié à reconnaitre les noms de personne. Tous les systèmes cités reposent sur une combinaison des expressions régulières dont leur configuration intègre une étape de filtrage. Selon les résultats expérimentaux, NERA 2.0 était

capable d'obtenir une amélioration par rapport au système NERA avec un F-mesure égal à 69,93%, 57,09% et de 54,28% pour les noms de personne, de lieu et d'organisation.

4.1.2. Système de Aboaoga et Aziz

Dans [Aboaoga et Aziz, 2013], les auteurs ont développé un système reconnaissant les EN ayant seulement la catégorie *nom de personne* dans des textes arabes. Le système proposé couvre trois domaines qui sont : le sport, la politique et l'économie et repose sur trois étapes dont la première étape est dédiée à effectuer un prétraitement. Le processus de prétraitement assure la tokenisation, le formatage des données et la décomposition en phrases d'un texte traité. La deuxième étape consiste à marquer automatiquement les EN en se basant sur une liste prédéfinie des noms de personne et de mots déclencheurs pour le processus d'annotation. La troisième étape est l'application des règles sur le texte afin de reconnaître des noms de personnes qui n'existent pas dans les dictionnaires créés. Le noyau du système développé est composé par quatre règles principales. Il faut mentionner que les auteurs déterminent une liste de mots déclencheurs obtenus grâce à l'analyseur morphologique BAMA (Buckwalter Arabic Morphological Analyzer) sachant que cet outil est développé par [Habash et al., 2009]. L'évaluation du système est faite sur un ensemble des données collectées à partir de différents journaux arabes en ligne tels que : koora.net, aleqt.net et Alquds.net. Les résultats empiriques montrent que le système atteint une F-mesure égale à 92,66% dans le domaine de sport plus élevées que celles obtenues dans les domaines politiques (92,04%) et économiques (90,43%). Le système proposé se concentrait que sur la catégorie nom de personne ce qui explique les valeurs obtenues lors de son évaluation. Cependant, la REN reste dans trois domaines restreints ce qui limite la liste des mots déclencheurs obtenus. En fait, la couverture de cette liste augmente en rajoutant de nouveaux domaines d'étude.

4.1.3. Système de Fehri

En se basant sur les transducteurs, dans [Fehri, 2012], l'auteur a proposé un module de REN pour la langue arabe dont le domaine choisi est le sport. Le module élaboré reconnaît les EN décrivant les *noms de lieux sportifs* pour faciliter leur traduction en français ultérieurement. La phase de REN repose sur une étude typologique approfondie permettant de distinguer les différentes catégories nécessaires et spécifiques au domaine choisi. Le module de REN proposé se compose de deux étapes dont la première est dédiée à identifier un ensemble de dictionnaires et des patrons syntaxiques à partir du corpus d'étude. En fait, l'auteur a transformé les patrons syntaxiques en des transducteurs à états finis ce qui constitue la deuxième étape. Les transducteurs conçus sont implémentés à travers la plateforme linguistique Nooj. Les résultats

obtenus ont été évalués à travers les métriques de performance ayant les valeurs 98% de précision, 70% de rappel et 82% de F-mesure. Le système proposé montre des valeurs encourageantes en revanche il est spécifique qu'au domaine du sport. L'application à un nouveau domaine ne se fait qu'à travers la découverte de nouveaux patrons syntaxiques. Par conséquent, le temps de mise à jour ou maintenance sera égal à celui d'établissement d'un nouveau système générique.

4.1.4. Système RENAM

Le système RENAM est le premier système développé par [Talha et al., 2014] permettant de détecter et extraire les EN pertinentes dans des textes amazighs. Ce système constitue un défi car il doit tenir compte les particularités associées à l'amazighe (langue berbère). Le système RENAM se base sur un ensemble de lexiques de noms propres et des règles construites manuellement en exploitant un outil d'extraction d'EN déjà disponible sous la plateforme libre GATE[2]. Le système RENAM se compose principalement de deux phases principales : la préparation du corpus et la REN. D'une part, la préparation du corpus consiste à collecter un recueil des données contenant des textes journalistiques amazighes. D'une autre part, cette préparation inclut la segmentation des textes en des phrases puis en des mots. La phase de REN s'effectue en créant des listes des noms propres et en développant manuellement des règles linguistiques (10, 4 et 5 règles dédiées respectivement pour les catégories *nom de personne*, *organisation* et *nom de lieux*). L'évaluation de résultats obtenus a été faite via les mesures de performance pour trois catégories. Pour la catégorie nom de personne, le système a obtenu 64%, 63% et 64% respectivement pour la précision, le rappel et la F-mesure. Ces valeurs ont été meilleures que celles obtenus pour la catégorie de nom de lieu qui a une valeur faible de précision qui égal à 27%, mais des valeurs acceptables 71% et 40% pour le rappel et la F-mesure. La reconnaissance des EN de la catégorie *organisation* a été la plus motivante en obtenant les valeurs 82%, 81% et 82% pour la précision, le rappel et la F-mesure. Le système RENAM a été capable de reconnaître juste trois catégories, un nombre réduit par rapport à d'autres travaux de REN. Il est recommandé au système RENAM d'avoir une phase d'analyse syntaxique plus profonde que celle effectuée afin de rajouter plus de critères de délimitation d'une EN étant donné que certaines EN complexes n'ont été pas reconnues correctement. Il faut remédier aussi le problème d'absence des informations morphologiques dans l'amazighe.

[2] Http ://gate.ac.uk

4.1.5. Système de Btoush et al

Un module de REN a été proposé par [Btoush et al., 2016] intégré dans un analyseur morpho-syntaxique (Part of Speech (PoS) tagging). Pour ce module de REN, les auteurs ont élaboré un détecteur reconnaissant des EN et agissant sur un texte donné en l'enrichissant avec des bonnes étiquettes attribuées aux EN reconnues. Le module de REN élaboré a la forme d'un détecteur ayant un ensemble d'instructions à suivre pour réaliser son objectif. Le détecteur proposé commence par la lecture d'un ensemble de données ayant la forme des listes qui stockent respectivement les noms de personne, de lieu et d'organisation. Ensuite, il segmente le texte d'entrée en des mots afin d'appliquer les règles construites pour la tâche de REN. Par conséquent, le détecteur effectue le test suivant : si les mots segmentés correspondent à une des règles proposées alors ce détecteur leur affecte l'étiquette associée, sinon ils seront considérés en tant que mots inconnus. Trois étiquettes ont été utilisées PERS, ORG et LOC pour décrire respectivement les catégories *nom de personne*, *nom de lieu* et ceux *organisation*. Le module de REN élaboré a été expérimenté sur un fichier contenant 490 mots et il était capable d'en reconnaître 480. Dans ce travail, les auteurs n'ont pas d'évaluation basée sur les mesures de performance pour les erreurs.

4.2. Systèmes statistiques de REN

Les systèmes de REN basés sur l'apprentissage supervisé utilisent des corpus annotés et prennent en paramètres des catégories qui sont la source de processus d'apprentissage. Par la suite, ils feront un appel à chaque catégorie passée en paramètres lors de la prise de décision sur celle adéquate à attribuer pour une EN. La manipulation des paramètres pour améliorer les résultats permet d'augmenter l'intelligence de ces systèmes. Les différents systèmes de ce type d'approche se basent notamment sur les méthodes d'apprentissage suivantes : Machines à Vecteurs de Support (SVM), modèle de Markov à Etats Cachés (HMM) [Bikel et al., 1997], modèle de l'Entropie Maximale (EM) [Borthwick et al., 1998], modèle de Champs Conditionnels Aléatoires (CRF) [Béchet et Charton, 2010], le Réseau de Neurone Artificiel (ANN) [Yegnanarayana, 1994] et les arbres de décision [Isozaki, 2001]. Dans ce qui suit, nous présentons quelques systèmes de REN qui se basent sur l'approche statistique classées selon les algorithmes d'apprentissages utilisés.

4.2.1. Système de Darwish et Gao

Pour améliorer la REN sur les micro-blogs, [Darwish et Gao, 2014] ont élaboré un système comportant trois méthodes : utilisation des dictionnaires larges extraits à partir de Wikipédia, un domaine d'adaptation et deux passages d'une méthode semi-supervisée qui se repose sur un

classificateur CRF. L'évaluation était faite sur ANERcorp et le nouvel ensemble d'entraînement de tweets (qu'ils étiquetaient) avec les mesures de performance égales à 76,8%, 56,6 % et 65,2% décrivant respectivement la précision, le rappel et la F-mesure. Le système élaboré annote les noms de personne, de lieu et d'organisation selon les catégories *PERS*, *LOC* et *ORG* et les EN qui n'appartiennent à aucune catégorie déjà citée par une balise qui s'appelle Overall.

4.2.2. Système de Kanya et Ravi

La technique CRF est utilisée par [Kanya et Ravi, 2016] pour proposer un système pour reconnaître ce genre d'EN en se basant sur un ensemble de documents pertinents extrait à partir de PUBMED qui est un catalogue bibliographique étendu du domaine biomédical. Le système proposé commence par la manipulation des documents sélectionnés à travers une phase de prétraitement composée de deux sous-étapes : la tokenisation ou dés-suffixation. Ce prétraitement constitue une phase importante dans ce système qui englobe aussi d'autres phases telles que l'annotation syntaxique. L'annotation sémantique est le cœur du système proposé utilisant des ressources lexicales et le modèle CRF. L'annotation se base sur une ontologie appelée Gene Ontology (GO) qui étiquettent les EN reconnues selon six classes qui sont : *gene*, *protein*, *DNA*, *RNA*, *Cell line* et *Cell type*. La recherche des corpus exploités est restreinte vu que les auteurs ont collecté trois classes de corpus seulement à travers des requêtes. Les trois classes sont *Breast cancer*, *Lung cancer* et *Thyroid cancer*. Le choix de ces trois classes a conduit à un nombre de catégories d'EN réduit.

4.2.3. Système de Salleh et al

Dans [Salleh et al., 2017], les auteurs ont proposé un système de NER pour la langue malaise qui se compose de trois phases principales. La première phase est dédiée au prétraitement qui est nécessaire pour la tokenisation et l'annotation de l'ensemble de données à travers un analyseur syntaxique en suivant un tagset. La deuxième phase concerne la génération du modèle de REN basé sur la technique CRF implémenté en Python. Dans la troisième phase, les auteurs évaluent leur système proposé à travers les mesures de performance pour quatre catégories qui sont *Facility*, *Person*, *Location* et *Organization*. Les résultats obtenus touchent une précision égale à 75%, un rappel égal à 72% et un F-mesure égale à 70%. Les résultats sont obtenus à cause d'au manque de ressources textuelles pour la langue malaise. Ces résultats peuvent s'améliorer en augmentant la taille du corpus et en variant les domaines d'étude pour toucher plusieurs formes d'EN. L'annotation des EN dans ce système souffre d'un nombre de catégories faible et d'exploration de corpus superficiel ce qui engendre un niveau de granularité réduit.

4.2.4. Système de Yao et al

La reconnaissance des EN biomédicales se fait également à travers la technique du réseau neuronal convolutif (CNN). Dans cette optique s'inscrit le travail de [Yao et al., 2015] dans lequel les auteurs proposent un système de REN basé sur une architecture de réseau de neurone approfondie. Cette architecture est composée de plusieurs couches dont chacune possède des caractéristiques abstraites dépendantes de celles générées par la couche inférieure. Les auteurs ont exploité des types de données : des données non annotées d'avance à partir de corpus GENIA et des données collectées à partir de la base de données PUBMED dont leur collection se fait cette fois à l'aide de l'outil Biopython[3]. Les auteurs assignent les EN biomédicales en cinq catégories qui sont : *protein*, *DNA*, *RNA*, *cell_type*, *cell_line*. Les résultats obtenus touchent une précision égale à 66,54%, un rappel égal à 76,13% et un F-mesure égale à 71,01%. La catégorie *gene* est manquante pourtant les auteurs ont exploité le PUBMED comme le travail illustré précédemment. Le travail présenté souffre d'un problème de reconnaissance d'une EN complexe. En outre, la reconnaissance en erreur du premier élément de l'EN engendre une reconnaissance incorrecte de la suite des éléments.

4.2.5. Système de Mohammed et Omar

En adoptant la technique ANN, [Mohammed et Omar, 2012] a tenté de résoudre les problèmes de la REN arabe. En fait, ils ont élaboré un système composé de trois étapes qui sont le prétraitement, conversion des lettres arabes en latin et la classification. La conversion des lettres arabes en latin vise à rendre les données en arabe plus accessible. Dans l'étape de classification, les auteurs ont fait recours le classifieur BPN (Back-Propagation Net) qui est le plus important parmi ceux associés à la technique d'apprentissage ANN. En fait, l'évaluation de ce système a montré qu'il atteint une précision égale à 92%. Son évaluation a été fait sur un corpus collecté à partir diverses sources Web et elle a été comparée aux arbres de décisions et obtient 87% de précision. Le système proposé ne reconnait que quatre catégories *nom de personne, nom de lieu, organisation* et *divers.*

4.3. Systèmes hybrides de REN

L'exploitation de l'approche hybride n'est pas récente vu que pour MUC-7 un système appelé LTG (Language Technology Group) a été élaboré par [Andrei et al., 1998]. Ce système permet de reconnaître les EN par une suite de règles qui se présentent sous la forme d'une liste de mots déclencheurs et d'un ensemble de patrons contextuels utilisant un étiquetage en parties du discours. Le passage des règles se fait à travers leur organisation pour effectuer une première

[3] http://biopython.org/wiki/Biopython

reconnaissance partielle. Par la suite, un algorithme probabiliste fondé sur la technique EM est utilisé pour étiqueter les EN. Le système élaboré inclut également une phase de désambiguïsation des EN. La phase d'étiquetage est effectuée à travers la norme d'annotation SGML pour marquer les trois catégories principales TIMEX, NUMEX et ENAMEX. Etant donné que le système était encore débutant dans cette approche, il a eu certains problèmes liés au traitement des termes majuscules et à l'analyse de contexte.

4.3.1. Système de Zribi et al

Le système de [Zribi et al., 2010] repose sur un ensemble de règles générées automatiquement pour extraire et classer les EN. Les auteurs ont proposé un autre ensemble de règles extraites manuellement pour corriger et améliorer les résultats obtenus. Le système est composé de trois phases principales : l'extraction de règles, leur validation et l'extraction des EN. Ces phases reposent toutes sur l'analyse morphologique lors de l'étape préparation qui dépend à son tour d'un lexique de nom propre. La 3ème phase inclut une étape de filtrage des EN déjà extraites à travers des règles de rectification. A titre d'indication, ce système a été évalué sur le corpus arabe appelé ANERcorp[4]. Il atteint des valeurs de F-mesure égalent à 80,28%, 86,42% et 64,15%pour les catégories *nom de personne*, *nom de lieu* et *organisation*. Le système proposé nécessite une amélioration au niveau des règles élaborées afin de prédire le contexte des EN.

4.3.2. Système de Küçük et Yazıcı

Pour la langue turque, [Küçük et Yazıcı, 2012] ont développé un système de REN ayant un module principal à base de règles. Ce module qui est conçu manuellement, contient les sources de connaissances pour des domaines spécifiques. Ce module principal fait recours à des dictionnaires des EN ayant les catégories suivantes : *nom de personne*, *nom de lieu* et *organisation*. Il repose aussi sur des patrons reconnaissant les catégories déjà mentionnées et les expressions numériques et temporelles. Le système enrichit la REN à base de règles en la transformant en un dispositif hybride qui apprend à partir des données annotées disponibles. Les auteurs n'ont pas exploité un corpus de références accessibles mais ils ont développé leur propre corpus d'apprentissage et d'évaluation.

4.3.3. Système de Shaalan et Oudah

Un système capable de reconnaître 11 catégories d'ENA (*personne, lieu, organisation, date, heure, prix, mesure, pourcentage, numéro de téléphone, numéro ISBN et noms du fichier*) a été proposé par [Shaalan et Oudah, 2014]. Le système possède un module de REN symbolique prenant le texte brut en en entrée et il lui applique les dictionnaires et les règles linguistiques

[4] http://www1.ccls.columbia.edu/~ybenajiba/downloads.html

pour avoir une version annotée en sortie. Ce dernier sera l'entrée d'un module de REN statistique dont lequel les auteurs effectuent nombreuses expérimentations en utilisant les arbres de décision, la technique SVM et les classificateurs de régression logistique pour évaluer sa performance. Son évaluation a été faite sur le corpus ANERcorp. Le système proposé atteint un F-mesure égal à 0,94 pour la catégorie nom personne, 0,90 pour les noms de lieu et 0,88 pour les noms d'organisation. La phase d'évaluation reste insuffisante vu qu'elle ne teste que les trois premières catégories malgré la variété de catégories étudiées.

4.3.4. Système de Hkiri et al

Le ANERcorp est utilisé également pour tester un système de REN pour la langue arabe élaboré par [Hkiri et al., 2016]. Pour réaliser la REN, le système proposé applique un modèle CRF, un lexique bilingue d'EN et des règles linguistiques spécifiques. Le système reconnaît trois catégories d'ENA qui sont *nom de personne*, *nom de lieu* et *nom d'organisation*. Cependant, le nombre de catégories reste insuffisant vis-à-vis le nombre de corpus exploités qui sont le ANERcorp, le corpus News Commentary et le corpus United Nations. Le système hybride a une performance sur le corpus News Commentary avec un F-mesure égale à 81,2% meilleure que sur le corpus ANER avec un F-mesure égale à 80.9%.

4.3.5. Système de Sharma et al

La REN est une tâche qui s'universalise, néanmoins elle reste encore un défi en langue indienne comme l'assamais. En réalité, l'assamais qui est une langue indo-européenne sous dotée sans ressources textuelles appropriées. De plus, les corpus qui sont disponibles ne sont pas nombreux en les comparant à ceux associés à d'autres langues comme le français et l'anglais. Pour cette raison, il existe quelques travaux de recherche visant à développer des systèmes performants pour la REN en assamais [Sharma et al., 2016a]. Dans ce contexte, le premier système hybride réalisant l'objectif déjà cité a été élaboré par [Sharma et al., 2016b]. Le système élaboré reconnaissant les EN assamais traite quatre catégories qui sont *nom de personne, nom de lieu, nom d'organisation* et o*ptionnel*. En fait, les auteurs utilisent trois étapes principales dont la première concerne l'approche statistique à travers l'utilisation des deux techniques d'apprentissage CRF et HMM. La deuxième étape est dédiée à l'approche symbolique consistant à utiliser un ensemble de règles construites manuellement via des listes prédéfinies stockant les noms de personne, de lieu et d'organisation. La dernière étape est l'annotation des EN reconnues.

4.3.6. Système de Ramesh et Sanampudi

Le domaine biomédical est présent aussi comme un centre d'intérêt des systèmes hybrides visant à améliorer la REN dans ce domaine. Dans le système de [Ramesh et Sanampudi, 2016], les auteurs ont élaboré un modèle hybride de REN dans des documents biomédicaux non structurés. Le modèle élaboré a comme tâche principale d'identifier les EN et les classifier dans cinq catégories *DNA*, *RNA*, *protein*, *cell-in* et *cell-type*. En fait, le modèle proposé combine les approches symbolique et statistique après une étape de prétraitement. L'approche symbolique est utilisée juste pour la phase d'identification tandis que leur classification est assurée par un classifieur SVM. Les auteurs ont exploité le corpus GENIA (Medline abstract collection) pour la phase d'expérimentation de leur modèle. Le système atteint 80,76% de précision, 84,24% de rappel et 80% de F-mesure. L'identification des EN se base sur un dictionnaire biomédical qui n'est pas exhaustif pour cette raison, les auteurs développent de nouvelles règles manuellement.

4.4. Discussions

Après avoir illustré des systèmes de REN reposant sur les trois approches fondamentales, nous constatons que le fonctionnement des systèmes symboliques s'appuie sur la construction des règles linguistique à travers une intervention humaine et manuelle. Dans certains cas, ces règles linguistiques construites subissent une révision par un expert linguistique. De plus, la conception de ces règles dépend des dictionnaires exploités lors de leur établissement. Partant de ce fait, la performance de ces règles s'améliore en augmentant la couverture des dictionnaires élaborés lors de la REN. Cependant, cette amélioration touche également les mots déclencheurs jouant le rôle des indices internes ou externes repérant les EN. Autrement dit, l'enrichissement de la liste des mots déclencheurs favorise l'augmentation du champ de reconnaissance. Il faut mentionner aussi que la maintenance d'un système symbolique pour ajuster ses fonctionnalités est difficile. En outre, la mise à jour d'une règle requiert la modification de toutes les règles qui se basent sur elle.

D'après les systèmes statistiques étudiés, nous remarquons que la reconnaissance et l'annotation des EN se fait sur des techniques qui requièrent une annotation préalable des corpus exploités. Cette annotation dépend du degré de supervision qui peut être fort, moyen ou faible. Au contraire du système symbolique, ce genre de système se maintient d'une façon automatique. En revanche, les systèmes statistiques souffrent du manque d'une forte catégorisation et leur balisage d'annotation est souvent défini selon des préférences et non selon des normes. La combinaison de deux approches symbolique et statistique a poussé les systèmes hybrides à exceller progressivement dans la tâche de REN en profitant de leurs avantages.

Tableau 3. Récapitulation sur les travaux de REN antérieurs

Auteurs	**Approche**	**Nombre de catégories**	**Formalisme/ technique**	**Type d'annotation**	**Langue**
Shaalan et Raza	Symbolique	1 (nom de personne)	Règles linguistiques	Balisage défini	Arabe
Shaalan et Raza	Symbolique	10	Règles linguistiques	Balisage défini	Arabe
Oudah et Shaalan	Symbolique	10	Règles linguistiques	Balisage défini	Arabe
Fehri	Symbolique	4	Transducteur	Balisage défini	Arabe
Aboaoga et Aziz	Symbolique	1 (nom de personne)	Expression régulière	Balisage de CoNLL	Arabe
Talha et al	Symbolique	3	Règles linguistiques	Balisage défini	Amazighe
Btoush et al	Symbolique	3	Règles linguistiques	Balisage de CoNLL	Arabe
Darwish et Gao	Statistique	4	CRF	Balisage de CoNLL	Arabe
kanya et Ravi	Statistique	6	CRF	GO	Anglais
Salleh et al	Statistique	4	CRF	Balisage défini	Malaise
Mohammed et Omar	Statistique	4	ANN	Balisage de CoNLL	Arabe
Yao et al	Statistique	5	CNN	Balisage défini	Anglais
Andrei et al	Hybride	3	Règles / EM	SGML	Anglais
Zribi et al	Hybride	4	Règles automatiques et manuelles	Balisage de CoNLL	Arabe
Küçük et Yazıcı	Hybride	-	-	Balisage défini	Turque
Shaalan et Oudah	Hybride	11	Arbre de décision/SVM / expression régulière	Balisage défini	Arabe
Hkiri et al	Hybride	3	CRF/ lexique bilingue/ Règles linguistiques	XML	Arabe
Sharma et al	Hybride	4	CRF/ HMM/ règles manuelles	Balisage défini	Assamais
Ramesh et Sanampudi	Hybride	5	Règles manuelles/ SVM	Balisage défini	Anglais

Le tableau 3 classifie les travaux de REN principalement selon l'approche et des facteurs de distinction. D'après ce tableau, les travaux de REN qui se basent sur les approches fondamentales (symbolique, statistique et hybride) ont un nombre de catégories exploitées réduit pour repérer les EN car elles n'atteignent pas une vingtaine de catégories dans les cas présentés. Cela peut être justifié au fait que ces travaux ne fixent pas d'avance une typologie d'EN étendue et raffinée. Pourtant, certains systèmes sont dédiés à un domaine particulier tel que le biomédical qui se réfère à des ontologies pour distinguer les catégories comme l'ontologie GO. Il existe aussi la variété de formalismes et techniques exploités qui participent à améliorer la REN. Cependant, nous trouvons que cette variété excelle en traitant une langue particulière comme la langue arabe. Le tableau 3 montre aussi que le type de de balisage diffère d'un travail à un autre mais ce balisage se fait souvent selon les balises des conférences CoNLL. La présence de techniques sophistiquées n'a pas influencé sur l'annotation dans certains travaux. En contrepartie, nous trouvons que [Andrei et al., 1998] s'appuie sur une norme, la norme SGML, bien que ces travaux soient anciens. Les auteurs n'adoptaient aucune définition formelle pour déterminer les limites d'une EN au cours de la REN et les systèmes élaborés souffrent d'un manque de ressources exhaustives. La plupart des systèmes élaborés choisissent un domaine particulier ce qui défavorise leur application sur d'autres types de corpus d'où vient l'absence d'uniformité. Les problèmes d'ambiguïté obtenus lors de la REN sont présents dans tous les travaux cela est à la cause d'un manque d'ordonnancement des tâches à réaliser au sein des systèmes élaborés. Pour cette raison, le choix d'un formalisme adéquat assurant l'ordre de passage des règles joue un rôle important dans la REN.

5. Relations sémantiques entre les entités nommées

L'extraction des RS est une étape importante qui consiste à identifier des liens sémantiques pertinents entre les EN. Pour cette raison, cette tâche est fortement liée à celle de la REN. En effet, extraire les RS dépend non seulement de leur définition précise, mais aussi de la catégorisation des EN. Cette catégorisation effectuée au cours de la REN permet de déterminer le sens des EN et de prédire les types des RS. Généralement, les systèmes d'extraction des RS se basent sur les approches fondamentales (symboliques, statistiques ou hybrides) comme les systèmes de REN. Nous divisons la section courante en quatre sous-sections dont la première est dédiée à la présentation d'un aperçu sur l'apparition de la tâche d'extraction des RS. Les sous-sections restantes sont consacrées aux systèmes d'extraction des RS élaborés en se basant sur les trois approches fondamentales.

5.1. Apparition de la notion de RS pour les EN

L'extraction des RS entre les EN a débuté comme une tâche indépendante lors de la conférence MUC-7 [Chinchor, 1998]. Cette tâche est formalisée pour structurer les données enrichissant les bases de données. Les RS traitées ont été considérées comme un lien significatif entre deux EN. Au cours de cette conférence, trois types ont été initialement proposés qui sont *empolye of*, *location of* et *product of* identifiés à partir d'un corpus journalistique. L'extraction des RS a été étendue pour qu'elle soit un objectif visé par d'autres conférences et campagnes d'évaluation suivant MUC-7 à savoir la campagne ACE [Doddington et al., 2004]. L'ACE a proposé une nouvelle définition pour la RS permettant sa délimitation en tant qu'un lien significatif entre les EN dans un texte et pouvant avoir de types principaux qui sont statiques ou événements. De plus, leur annotation a été orientée par un guide. Actuellement, la tâche d'extraction des RS fait l'objet d'une série de campagne d'évaluation internationale appelée TAC[5] (Text Analysis Conference) -KBP (Knowledge Base Population) qui a été lancé en 2008 jusqu'à présent. L'objectif TAC s'agit d'encourager les systèmes à reconnaître une soixantaine de relations sémantiques reliant différentes sortes d'EN (essentiellement des lieux, des personnes, des organisations et leurs différentes sous-catégories). Après leur élaboration, ces systèmes passent par une phase de validation de l'ensemble de RS fournies.

Nous constatons que les définitions attribuées au RS peuvent avoir un élargissement pour qu'elles soient précises et ciblées. Cet élargissement permet non seulement leur délimitation mais déterminer leurs types également. Dans ce qui suit, nous commençons par présenter les systèmes d'extraction des RS symboliques.

5.2. Systèmes symboliques d'extraction de RS

Rappelons que les systèmes symboliques sont caractérisés par la création des grammaires locales et formelles décrites par des règles construites manuellement. Ces règles peuvent être modélisées à travers le formalisme adéquat comme par exemple les expressions régulières et les transducteurs. Dans cette section, nous présentons les systèmes d'extraction de RS basées sur des règles et ceux fondés à la base des transducteurs.

5.2.1. Système de Ben abacha et al

Pour le domaine médical, [Ben abacha et al., 2011] ont proposé un système d'extraction des RS entre les EN médicales. Ce système repose sur une étape de prétraitement consistant à reconnaître et catégoriser les EN médicales. Les RS ont été extraites à partir de chaque couple d'EN sachant que la détermination de leurs types s'effectue à travers un réseau sémantique

[5] https://tac.nist.gov/

appelé UMLS (Unified Medical Language System). A la base d'UMLS, le système a réussi d'extraire les six types de RS suivants : *causes*, *diagnoses*, *treats*, *prevents*, *complicates*, *sign* et *sympton of*. Les auteurs ont implémenté leur système sur la plateforme MeTAE (Medical Texts Annotation and Exploration). En fait, le processus d'annotation pour les RS et les EN est effectué en RDF (Resource Description Framework). Finalement, les auteurs ont expérimenté leur système sur un corpus collecté à partir de PCM (Pub-MED Central).

5.2.2. Système IMAIOS

Dans [Lafourcade et Ramadier, 2016], les auteurs ont proposé un système appelé « IMAIOS » dont sa tâche principale est d'extraire des RS à partir d'un ensemble de textes non structurés. En fait, cet ensemble de textes forme un corpus qui a été collecté à partir des rapports radiologiques français. Dans IMAIOS, les auteurs ont construit des patrons linguistiques associés à des contraintes sémantiques. Le choix des contraintes n'est pas arbitraire de sorte qu'il soit vérifié à travers un réseau lexico-sémantique appelé JDM (Jeux De Mots). Réellement, le système proposé est capable d'extraire 15 types de RS et de les identifier en respectant les consignes des radiologistes.

5.2.3. Système de Ghamnia

Afin d'enrichir les pages de désambiguïsation de la ressource libre DBpédia à partir de Wikipédia, [Ghamnia, 2016] a proposé un extracteur de relations d'hyperonymie dédié aux types de pages déjà mentionnées. L'extracteur est basé sur des patrons lexico-syntaxiques qui ont été conçus et développés sous formes d'expressions régulières. Pour réaliser cet extracteur, l'auteur a constitué un corpus à partir du téléchargement de la Wikipédia française. Ensuite, il a effectué une phase de nettoyage de ce corpus pour extraire le texte à partir de la version XML. Le corpus nettoyé était lui-même l'entrée d'une phase d'étiquetage morpho-syntaxique en utilisant l'outil TreeTagger[6]. L'auteur a fait recours à un extracteur de termes appelé YaTeA[7] pour identifier les syntagmes nominaux se trouvant dans une relation d'hyperonymie. Puis, il a transformé les différentes formes identifiées pour ce type de relation en des patrons lexico-syntaxiques exécutables, sous forme d'expressions régulières, sur la plateforme Gate.

5.2.4. Système de Ezzat

[Ezzat, 2014] a proposé une méthode semi-automatique pour obtenir des grammaires locales qui extraient des RS entre les EN dans les corpus. La tâche de détection d'EN est assurée par un analyseur appelé Arisem basé également sur l'approche symbolique. Cet analyseur annote

[6] http ://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger
[7] https://perso.limsi.fr/hamon/YaTeA/?lang=fr

les mots et les syntagmes pertinents en utilisant des étiquettes sur différents niveaux du texte (morphologique, syntaxique et sémantique). La méthode proposée extrait une seule relation appelée « *contact* ». Pour annoter cette relation, l'auteur utilise XML (eXtensible Markup Language) tandis que l'évaluation de méthode a été faite sur un échantillon d'articles de journaux (Le Monde 2007).

5.2.5. Système de Ben Hamadou et al

Dans [Ben Hamadou et al., 2011], les auteurs ont proposé un système pour extraire les RS entre les EN. Les RS étudiées sont de type fonctionnel (*Directeur*, *Responsable* et *Président*) qui relie les catégories Personne et Organisation. Le système est un processus composé par trois étapes dont la première est dédiée à la REN, plus précisément celle des deux catégories déjà mentionnées. La seconde étape s'occupe de l'extraction des RS entre les EN. La dernière étape est consacrée à la génération de prédicats représentant les formes associées au type de RS identifiée. En fait le système proposé permet également la traduction des RS détectées en français. Les auteurs ont implémenté leur système sur la plateforme linguistique Nooj alors que l'expérimentation a été réalisée en utilisant un corpus journalistique et un autre corpus extrait à partir de la Wikipédia. Cependant, l'annotation utilisée ne respecte aucune norme.

5.3. Systèmes statistiques d'extraction de RS

L'approche statistique utilise des algorithmes d'apprentissage pour apprendre les décisions de balisage des RS à partir des corpus annotés. Cette approche exige la présence d'un nombre important de données annotées. Dans ce qui suit, nous présentons quelques systèmes basés sur des règles générées automatiquement et ceux basés sur des techniques d'apprentissage.

5.3.1. Système de Abd El-Salam et al

[Abd El-Salam et al., 2016] ont proposé un système semi-supervisé permettant l'extraction des relations binaires entre les EN. L'extraction s'effectue en se basant sur un ensemble de textes arabes collectés à partir du contenu libre du Web. Les auteurs considèrent que leur système est générique et peut être intégré sur différents domaines. En fait, ce système est décrit par un processus itératif d'extraction des RS dont chaque itération contient principalement l'extraction des patrons et celle des instances. Dans chaque itération, les nouvelles instances subissent un filtrage évitant les bruits. Le système proposé extrait quatre types de RS qui sont : *author-of* (personne, équipe), *president-of* (personne, pays), *play-in* (personne, équipe) et *CEO-of* (personne, société). Les auteurs n'ont pas respecté une norme lors de l'annotation de ses RS.

5.3.2. Système de Tourille et al

L'extraction des informations temporelles à partir des enregistrements électroniques de la santé est devenue un centre d'intérêt important. La nécessité d'un processus d'extraction est justifiée par le fait que chaque équipe médicale a toujours besoin d'accéder à des informations à partir de perspectives temporelles. Dans ce contexte, nous citons le travail de [Tourille et al., 2017] qui consiste à proposer un modèle pour extraire des relations temporelles à partir des récits cliniques en français et en anglais. Le modèle proposé est basé sur la technique des SVM dont son élaboration se compose de deux tâches principales. La première tâche s'appelle DR (Document creation time Relation) ayant comme objectif la localisation évènements médicaux (EVENT) existant dans des documents DCT (Document Creation Time). Dans cette tâche, les auteurs utilisent les balises d'annotation suivantes : *Before*, *Before-overlap*, *Overlap* et *After*. La deuxième tâche composant le modèle s'appelle CR (Container Relation) qui permet d'identifier les relations temporelles, exprimant une inclusion, entre des paires d'entités (EVENT et TIMEX). L'expérimentation a été effectuée sur deux corpus comparables qui sont respectivement le MERLOT[8] (Multimedia Educational Resource for Learning and Online Teaching) pour la langue française et le THYME[9] (Temporal History of Your Medical Events) pour celle anglaise.

5.4. Systèmes hybrides d'extraction de RS

Les systèmes hybrides dédiés à l'extraction des RS intègrent deux modules dont le premier est généralement basé sur l'approche symbolique pour étiqueter la ressource textuelle exploitée. Étant donné que cette ressource est annotée, le système applique son deuxième module statistique afin de ré-étiqueter les RS détectées dont l'objectif est d'améliorer la performance de ce système. Dans ce contexte, nous présentons quelques systèmes hybrides.

5.4.1. Système de Lahbib et al

Pour la langue arabe, [Lahbib et al., 2013] ont développé un système d'extraire des RS arabes dont l'idée principale est d'exploiter des dépendances syntaxiques pour inférer les RS. Ce système combine des calculs statistiques et des connaissances linguistiques. En fait, deux étapes composent le système proposé sachant que la première étape consiste à utiliser un outil linguistique. Cet outil utilisé permet d'analyser les textes et d'extraire trois types de RS. Dans la deuxième étape, les auteurs font recours à des mesures statistiques pour compter la similarité entre les EN reliées. En effet, ils préfèrent aussi d'exploiter des textes voyellés. Pour cette

[8] https://www.merlot.org/merlot/index.htm
[9] https://github.com/stylerw/thymedata

raison, le système est expérimenté sur le corpus contenant des textes représentant les hadiths, plus précisément une version voyellée et structurée.

5.4.2. Système de Boujelben et al

Un autre système traitant la langue arabe a été proposé aussi par [Boujelben et al., 2014] dont le module linguistique a été utilisé comme une deuxième étape au contraire du système déjà décrit. En fait, les auteurs ont proposé un système pour extraire des RS entre les ENA. Initialement, les résultats ont été obtenus à travers un module statistique. Le module linguistique a joué le rôle un post-traitement pour améliorer les résultats déjà obtenus. Le système élaboré extrait des RS simples et complexes exprimées par un ou des ensembles de mots. Le processus d'annotation est fait avec un balisage défini qui ne respecte aucune norme.

5.5. Discussions

Les systèmes d'extraction des RS que nous avons présentés reposent sur différents formalismes et techniques selon l'approche exploitée. La différence entre eux se présente au niveau du nombre de types de RS extraits et au niveau du balisage utilisé pour leur annotation. Dans le tableau suivant, nous proposons une récapitulation des travaux déjà illustrés classifiées selon facteurs de distinction.

Tableau 4. Récapitulation sur les travaux antérieurs d'extraction des RS

Auteurs	Approche	Formalisme/ technique	Nombre de types de RS	Type d'annotation	Langue
Ezzat	Symbolique	Transducteurs	1 (contact)	XML	Français
Ben Hamadou et al	Symbolique	Transducteurs	1 (Fonctionnelle)	XML	Arabe
Ben abacha et al	Symbolique	Règles linguistiques	6	RDF	Anglais
Lafocarde et Ramadier	Symbolique	Patrons linguistiques	15	Balisage défini	Français
Ghamnia	Symbolique	Expressions régulières	1 (hyperonymie)	XML	Anglais
Abd El-salam et al	Statistique	Technique semi-supervisé	4	Balisage défini	Arabe
Tourille at al	Hybride	SVM	1 (Relation temporelle)	Balisage défini	Français/ Anglais
lahbib et al	Hybride	Règles linguistiques/ Mesures statistiques	-	Balisage Défini	Arabe
Boujelben et al	Hybride	Transducteurs/ Arbre de décision	14	XML	Arabe

Le tableau 4 résume les travaux d'extraction des RS étudiées qui diffèrent selon le nombre de RS qui varie d'un domaine à un autre. Il faut mentionner que les travaux étudiés n'ont adopté aucune définition formelle et précise d'une RS. En outre, ils n'indiquent pas non plus de références aux définitions déjà proposées par les conférences et les campagnes d'évaluation. Concernant l'annotation, nous avons remarqué que ce processus utilise généralement un balisage bien défini et associé aux préférences des auteurs quoiqu'il existe aussi des systèmes exploitant des normes d'annotation comme RDF et XML. D'autre part, nous constatons que tous les systèmes avaient une étape de prétraitement incluant un processus de REN. Le processus de REN facilite l'identification des liens sémantiques entre les EN reconnues. Il faut mentionner aussi que certains systèmes créent leur propre module de REN dépendant ou non de l'approche utilisée et d'autres font recours à des outils existants à savoir l'analyseur Arisem et MetaMap. Les EN reconnues dans ces systèmes ne sont pas affectées à un nombre important de catégories ce qui a influé sur le nombre de types de RS détectées. D'ailleurs, ce nombre réduit de catégories est dû également à la nature des corpus exploités car la majorité de ces corpus appartient à des domaines restreints.

Revenons au processus de REN qui précède toujours la tâche d'extraction des RS. En améliorant la précision de la REN, la performance de la tâche d'extraction des RS augmente. Dans ce cas, une catégorisation approfondie est toujours nécessaire afin d'augmenter le niveau de granularité fournissant un nombre important de types de RS entre les ENA.

6. Wikipédia

Les ressources libres sont des recueils de données informatisées de productions langagières écrites ou parlées. Ces ressources sont généralement utilisées pour l'enrichissement d'un contenu textuel. Elles fournissent l'exploitation automatique des informations linguistiques. Il faut assurer alors la qualité et la cohérence de ces informations afin de permettre la bonne exploitation. Parmi les ressources libres, nous citons la Wikipédia qui est une encyclopédie multilingue créée par Jimmy Wales et Larry Sanger le 15 janvier 2001. Cette encyclopédie apporte des liens explicatifs et offre un contenu librement réutilisable, objectif et vérifiable. Elle fournit aussi l'accessibilité et la reconnaissance automatique des sujets mentionnés dans des textes non structurés à travers des liens.

6.1. Volume arabe de la Wikipédia

Pour le volume arabe, les articles de la Wikipédia sont généralement écrits en ASM (Arabe Standard Moderne), parfois en AC (Arabe Classique) et AD (Arabe Dialectal). Rappelons qu'ASM, AC et AD sont trois styles de la langue arabe [Alsayadi et ElKorany, 2016].

Réellement, la diversité de ces styles dépend du contributeur qui a rédigé ou mis à jour les articles. En fait, la Wikipédia arabe est une ressource textuelle accessible, libre et riche en termes d'ENA et de RS, ce qui favorise leur traitement. De plus, la Wikipédia possède des zones d'informations pertinentes comme les infoboxes.

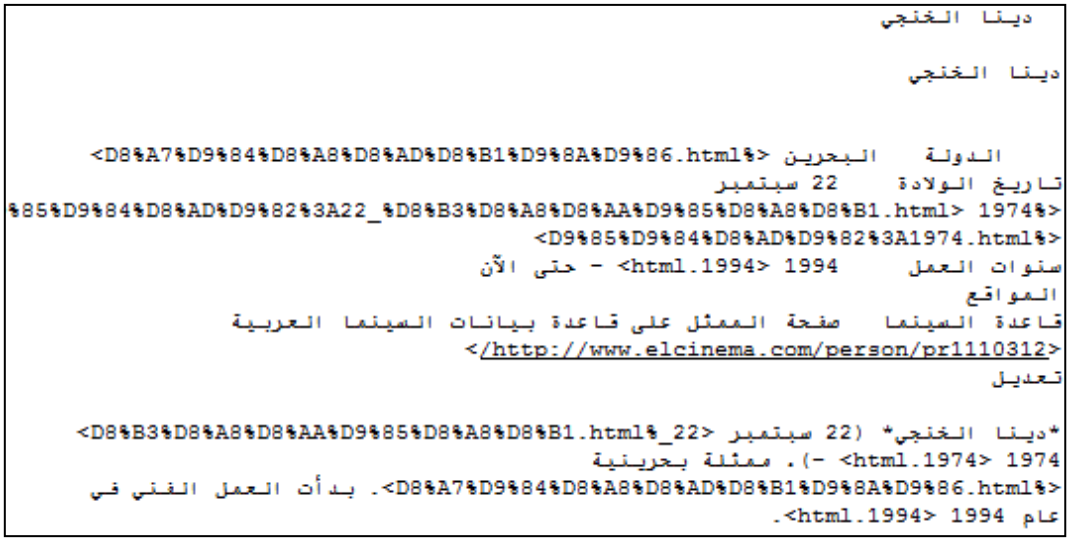


Figure 1. Article de la Wikipédia arabe et sa structure interne

La figure 1 décrit un article faisant partie de l'article de la Wikipédia arabe avec sa structure interne affichée sous un éditeur de texte. La récupération s'effectue via un outil appelé Kiwix permettant d'avoir une version hors ligne de cette ressource libre. Cet outil permet de faciliter la visualisation et l'analyse des articles sélectionnés pour effectuer un traitement. La variation de la qualité des articles Wikipédia permet de traiter toutes les formes alternatives qd'une EN arabe. En fait, une EN arabe peut apparaître dans divers contextes tels que les titres d'article et de section et les liens de redirection, soit vers la page appropriée, soit vers des mots clés correspondant à un pays particulier.

6.2. Travaux exploitant la Wikipédia

Parmi les domaines qui ont choisi la ressource libre Wikipédia pour élaborer des systèmes puissants, nous citons celui de la traduction automatique. Dans ce contexte, [Sellami at al., 2013] ont proposé un système de traduction automatique statistique à partir de corpus comparables. Dans ce système, les auteurs ont choisi la paire de langues arabe-français. Le travail effectué consiste à exploiter les liens inter-langues qui relient les articles en arabe à ceux en français. Ce type de lien facilite l'extraction entre les termes (simples ou composés) arabes et leurs traductions en français et vice versa. Plusieurs travaux s'inspiraient de cette ressource pour réaliser des enrichissements de leurs systèmes. En fait, la ressource Wikipédia est très utilisée pour construire des ressources linguistiques et les enrichir également. C'est dans ce

cadre que s'inscrit le travail de [Sellami et al., 2012] consistant à construire un lexique bilingue à partir de la Wikipédia en profitant de son aspect multilingue.

La richesse de la Wikipédia en termes d'EN et son aspect multilingue ont joué un rôle important pour la proposition des systèmes de REN. Parmi ces travaux, nous citons le travail de [Biltawi et al., 2016] qui se concentre sur la création des dictionnaires en exploitant cette ressource libre. Les dictionnaires créés sont classés en trois principales catégories : *nom de personne, nom de lieu et nom d'organisation*. Cette création de différents types de dictionnaires est une initiation à un processus de REN.

Les articles composants les ressources linguistiques libres sont très exploitées par des plateformes linguistiques pour leur manipulation, analyse et annotation. Dans le but d'annotation, ces articles subissent des processus de structuration basés sur différentes manières comme l'annotation basée sur un balisage défini. Ce type d'annotation nécessite l'appui sur norme d'annotation pour avoir une sortie structurée capable de s'intégrer dans le Web. Pour cette raison, nous allons présenter quelques normes d'annotation exploitées pour la manipulation des ressources textuelles.

7. Norme d'annotation TEI

Le processus d'annotation consiste à rajouter des informations interprétatives à des documents (textuel, auditif et visuel). Il existe trois types d'annotation dont le premier type peut être sous la forme d'une addition (manuelle) de certaines remarques ou des commentaires au contenu d'un texte. L'annotation peut se faire également à travers l'ajout des métadonnées dans les documents ou encore les corpus pour les rendre numérisés. Le dernier type est l'annotation linguistique qui touche à son tour les trois niveaux suivants : morphologique, syntaxique et sémantique. La manière d'annotation peut être manuelle via une interprétation humaine ou automatique via des outils [Eshkol-Taravella, 2015].

Diverses normes et conventions d'annotation (comme par exemple Quaero) sont proposées dans la littérature à travers lesquelles le processus d'annotation permet la génération ou l'échange des documents compréhensibles et elles assurent la portabilité de ces documents échangés. Le XML[10] et la TEI[11] sont parmi les normes représentant le contenu d'un document à travers des balises et des éléments riches. Le XML est une norme dérivée du standard SGML (ISO 8879). Cette norme joue un rôle signifiant à l'échange d'une variété de données dans le

[10] https//www.w3.org.XML/
[11] 5 https//www.tei-c.org

Web. Dans ce qui suit, nous allons nous donnons un aperçu sur la norme TEI et présenter les travaux qui l'exploitent plus précisément pour la langue arabe.

7.1. Aperçu sur la TEI

La norme TEI est un projet international visant à préparer des balises standards pour des documents ayant un format électronique [Text Encoding Initiative Consortium, 2018]. L'objectif est donc de faciliter leur échange. Ces documents peuvent être des textes, des images et des vidéos. La recherche sur cette norme a mené à la cinquième version appelée TEI P5. La TEI s'intéresse à l'encodage des métadonnées d'un document à travers la description de ses composants (paragraphes, phrases, etc.) afin de détailler sa structure à l'aide de divers types de balises. Cette norme offre aussi la possibilité d'avoir une analyse sémantique qui s'intéresse à la signification du mot.

La TEI conserve les deux types de structure d'un document qui sont physique et logique. La structure physique est une description simple du document qui peut englober le titre, le corps du texte et l'annexe. La structure logique organise le contenu d'un texte en le découpant en des parties, des chapitres et des sections. A ces deux types déjà mentionnés, la TEI ajoute la notion de métadonnées encodé dans une balise appelé <header> qui se place avant le corps du texte. La TEI offre un type de balisage appelé onomastique qui met en évidence les noms propres (par exemple les noms de personne et les toponymes) et les expressions temporelles comme les dates. Pour les toponymes, ils sont encodés grâce à la balise <placeName> qui correspond à désigner les noms de lieu absolus et relatifs (village, ville, région, pays, etc.). La date calendaire s'encode en TEI à travers la balise <date> pouvant contenir n'importe quel format d'une date. Au sein des balises TEI, certains éléments peuvent s'ajouter comme @type qui spécifie le type d'une information encodée [Dufournaud et al., 2012].

7.2. Travaux exploitant la TEI pour la langue arabe

Dans [Soualah et Hassoun, 2012], les auteurs ont proposé une adaptation de la TEI version P5 à la description des manuscrits en langue arabe. Cela permettra d'avoir des nouveaux manuscrits normalisés lisibles. La TEI a été utilisée à résoudre certains aspects complexes liées à l'arabe comme la présentation de la translitération d'un mot.

Dans [Maraoui et Haddar, 2015], les auteurs ont élaboré un prototype d'automatisation de l'encodage des lexiques arabes en TEI. Ce prototype contient quelques rectifications par rapport au modèle du modèle TEI ordinaire en considérant les spécificités de la langue arabe. Les auteurs ont envisagé le défi de l'automatisation des bases de données lexicales arabe via la norme TEI qui devient une nécessité difficile à réaliser. Le fait de les automatiser revient à créer

une structure descriptive qui couvre toutes les propriétés et les caractéristiques morphologiques, syntaxiques et sémantiques des entrées lexicales stockées dedans.

Dans [Lhioui et al., 2016], les auteurs ont exploité une base de données lexicale en TEI pour proposer une méthode d'interopérabilité entre des ressources lexicales ayant divers formats. La manipulation de ce genre de ressources n'est pas une tâche facile telle que leur fusion, la comparaison ou la recherche d'une correspondance entre eux. D'ailleurs, assurer l'interopérabilité entre les ressources lexicales a été considéré comme une tâche irréalisable.

[Ben Ismail et al., 2017] ont proposé un éditeur lexical arabe appelé ALIF associé à un vérificateur de contraintes basé sur les deux standards d'ISO (International Organization for Standardization) LMF (Lexcial Markup Framework) et la TEI. Les auteurs ont utilisé également l'éditeur appelé Oxygen pour créer les fichiers de règles associées à la TEI et vérifier leurs structures. Ce travail réalisé vise essentiellement à effectuer La normalisation des ressources lexicales arabes qui est une piste de recherche en cours d'amélioration.

La TEI a été exploité par [Maraoui et al., 2018] pour être la base d'un outil de segmentation d'un corpus extrait à partir de Sahih Albukhari. Cette segmentation exploite les balises TEI pour annoter les phrases et les syntagmes délimités.

L'annotation des documents textuels via des normes peut répondre aux besoins des applications à établir et faciliter leur exploitation ultérieurement. Cependant, ce n'est pas possible de confirmer toutes les balises et ses éléments associés à une norme par rapport aux phénomènes linguistiques d'une langue traitée. Certaines balises ne peuvent pas être prises en compte définie et d'autres balises seront adaptées selon la convention d'annotation.

8. Conclusion

Dans ce chapitre, nous avons exploré des travaux dédiés à la reconnaissance des EN en plusieurs langues ainsi que l'extraction des RS entre elles. La présentation des EN nous a permis de distinguer deux sortes de définitions ; à travers une définition proposée ou à travers les différentes catégorisations faites sur les EN. La définition implicite participe non seulement à délimiter les EN mais à les structurer également à travers des catégories et des sous-catégories. Elle permet aussi de faciliter leur annotation ultérieurement. Nous avons présenté les approches fondamentales d'EI en illustrant leurs principes et la différence entre elles. Ces approches sont la base de l'établissement des systèmes de REN et ceux d'extraction des RS car elles fournissent divers formalismes et techniques pour réaliser leur objectif. La présentation de ces systèmes nous a montré les résultats obtenus par chaque approche utilisée selon la langue traitée et la nature de corpus. Cette présentation nous a montré les lacunes posées par chaque système et elle a prouvé que l'étude d'un corpus pour un domaine particulier est moins compliquée par

rapport à l'indépendance d'un domaine. En illustrant les systèmes d'extraction des RS, nous avons trouvé qu'ils englobent des modules de REN internes ou ils font l'appel à des systèmes de REN existants. Il était clair que le nombre catégories issue de la REN a une importance pour prédire les types de RS. Nous avons consacré une partie pour décrire la Wikipédia plus précisément son module arabe qui se caractérise par la variété de ses styles d'écriture. Cette ressource présente plusieurs domaines d'étude ce qui favorise son exploitation dans les applications de REN ou d'extraction des RS. Nous avons conclu ce chapitre par les normes d'annotation assurant la compréhension et l'échange des documents textuels et elle garantit également leur interopérabilité. Nous nous sommes concentrés sur la norme TEI qui devient de plus en plus utilisée pour l'annotation des textes arabes.

Dans le chapitre suivant, nous allons donner un aperçu sur le formalisme automate et transducteur qui est exploité dans l'approche symbolique pour l'extraction d'information.

Chapitre 2 : Aperçu sur les automates et les transducteurs

Dans le domaine d'EI, divers formalismes (automates à nombre fini d'états, les transducteurs et les réseaux de transition) sont exploités par les chercheurs afin d'aboutir des modèles formels définissant rigoureusement leurs méthodes proposées. Cependant, ces formalismes peuvent ne pas répondre exactement à leurs besoins au niveau d'analyse lexicale des données traitées. L'analyse morphologique est parfois non abordée par ces formalismes lors du traitement des phénomènes linguistiques associés aux langues ayant une morphologique complexe. Au niveau d'analyse sémantique, les travaux de recherches existants se bloquent car ils n'ont pas abouti à un compromis au sein d'un formalisme exploité sur le couplage de l'extraction des informations à partir d'un texte et leurs annotations. La protection des informations extraites et annotées reste encore une responsabilité non prise en compte par certains formalismes. Ces informations doivent être protégées pour subir d'autres traitements pouvant engendrer des ambiguïtés. L'ordonnancement des instructions réalisées par le formalisme choisi devient un défi car son absence rend la tâche à réaliser manque de précision et souffre d'un taux d'erreurs élevé. En fait, l'existence de l'ordonnancement reste insuffisante jusqu'à fixer un ordre précis défini pour éviter l'ambiguïté. La manipulation des informations à extraire via un formalisme nécessite aussi une plateforme linguistique assurant son fonctionnement et son adaptation aux besoins d'applications à réaliser. La plateforme linguistique doit fournir un bon enchainement des tâches et fournir des nouveautés par rapport à celles fonctionnant d'une façon classique.

Le présent chapitre se compose de sept parties dont la première est dédiée à donner un aperçu sur le formalisme d'automate. Cette partie inclut la présentation du principe de fonctionnement d'un automate et ses différents types associés. Dans la deuxième partie, nous abordons l'automate à nombre fini d'états à travers sa présentation formelle et l'explication des opérations qu'il peut subir avec des exemples. Dans la troisième partie, nous expliquons les réseaux de transition et les types qu'ils peuvent avoir via des exemples. La présentation des transducteurs à nombre fini d'état se fait dans la partie quatre via sa représentation formelle et celle graphique qui est dépendante et indépendante d'une plateforme linguistique. Nous donnons aussi les domaines d'application des transducteurs à nombre fini d'états en s'appuyant sur des exemples explicatifs. Dans la cinquième partie, nous rappelons le principe d'une cascade de transducteurs et les applications qui l'exploitent dans divers domaines vont être illustrées dans la partie six. La dernière partie s'intéresse à la présentation de la plateforme linguistique Unitex qui est dédiée au TAL et qui manipulent les transducteurs à nombre fini d'états. Dans cette partie, nous nous effectuons une comparaison entre Unitex et une autre plateforme via divers critères.

1. Automates

Un automate est un modèle abstrait qui n'a pas une existence physique. Ce modèle permet de traiter l'information à travers la transformation d'un problème donné à un langage. Notamment, l'automate essaie d'analyser chaque élément de ce langage transformé pour résoudre le problème posé. Plusieurs applications du TAL reposent sur la notion d'automate parmi lesquelles nous citons l'analyse lexicale, morphologique et syntaxique et la REN plus précisément les systèmes basés sur l'approche symbolique [Fehri, 2012 ; Mesfar, 2008].

Le principe de fonctionnement d'un automate est décrit comme suit : étant donné que le problème est découpé en instances, chaque instance est représentée par un mot. Par la suite, un test sera effectué consistant à voir si ce mot appartient au langage représentant ce problème ou non. En se basant sur le résultat positif ou négatif de ce test, l'automate décide si le problème admet une solution ou non. Schématiquement, un automate est caractérisé par un ensemble de variables discrètes d'entrée, de sortie et d'états internes.

Dans ce travail, nous nous intéressons à plusieurs types d'automates : les automates à nombre fini d'états, les réseaux de transition et les transducteurs. La représentation de chaque type est effectuée de divers diagrammes de transitions. En fait, il existe une équivalence entre certains types. Dans ce qui suit, nous commençons par l'automate à nombre fini d'états.

2. Automate à nombre fini d'états

Un automate à nombre fini d'états est une machine abstraite qui reconnaît l'appartenance ou la non appartenance d'un mot à un langage régulier donné. Cette machine est un modèle théorique de référence dont ses constituants sont un alphabet, un ensemble d'états et une relation de transitions entre ces états.

2.1. Représentation formelle d'un automate à nombre fini d'états

Formellement, un automate à nombre fini d'états ayant un alphabet $\sum$ est un quintuple $(Q, \sum, q_0, F, \delta)$ tel que :

- Q est un ensemble fini (et non vide) appelé ensemble des états de l'automate,
- $\sum$ est un alphabet (un ensemble fini et non vide de symboles),
- q_0 est l'état initial ou encore état de départ et un élément de Q,
- F est un sous-ensemble de Q appelé l'ensemble des états finaux,
- δ est une fonction de transitions.

D'après la définition formelle d'un automate à états finis, nous constatons que celui-ci admet une configuration selon les différents éléments. Cette configuration est une paire représentée

par un état et un mot tel que (q, w) où q ∈ Q et w un mot de l'alphabet ∑. Dans la figure suivante, nous illustrons le principe à travers un exemple explicatif.

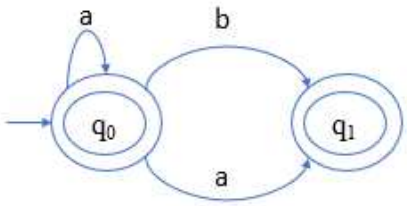


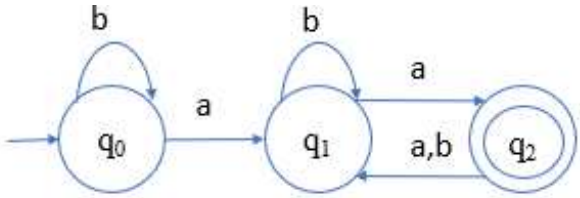
Figure 2. Diagramme de transition d'automate à nombre fini d'états

La figure 2 montre un diagramme de transition d'un automate à nombre fini d'états qui possède deux états q_0 et q_1. L'état q_0 appartient à Q et F. Cet automate possède un autre état final qui q_1 appartenant à F. À partir de schéma illustré, nous constatons qu'il existe plusieurs configurations possibles. Prenons quelques exemples de ces configurations : entre les états q_0 et q_1 nous avons $\delta(q_0, b) = q_1$, $\delta(q_0, ba) \rightarrow \delta(q_1, a) \rightarrow \delta(q_0, \varepsilon)$ et $\delta(q_0, a) = q_0$.

2.2. Automate à nombre fini d'états déterministe ou non déterministe

Etant donné un automate non déterministe A, il existe un automate déterministe A' qui reconnaît le même langage [Straubing et Weil, 2012]. Néanmoins, la différence entre les deux automates est que les automates finis déterministes possèdent un seul état initial I= $\{q_0\}$. Autrement dit, l'automate devient A= $\{Q, \Sigma, \delta, q_0, F\}$et $\forall q$ et a, Card(δ (q, a)) ≤ 1.

Soit l'automate A fini déterministe :



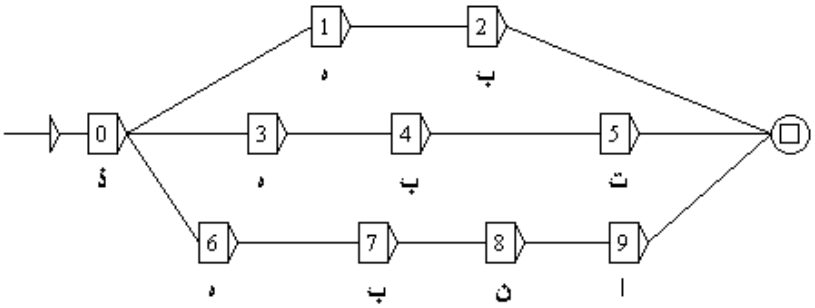
- Q= $\{q_0, q_1, q_2\}$ est un ensemble fini non vide d'états ;
- Σ= {a, b} est un alphabet fini de symboles ;
- F= $\{q_2\}$ est une relation de transition étiquetée qui relie deux états ;
- q_0 est l'état initial ;
- La relation de transition est la suivante :

$\delta(q_0, a) = q_1$, $\delta(q_0, b) = q_0$, $\delta(q_1, a) = q_2$, $\delta(q_1, b) = q_1$, $\delta(q_2, a) = q_1$ et $\delta(q_2, b) = q_1$.

Un automate fini A = (Σ, Q, δ, I, F) est dit non déterministe s'il existe dans δ deux transitions (q_1, x, q_2) et (q_1, x, q_3) telles que $q_2 \neq q_3$. C'est-à-dire, dans le même automate, il existe plusieurs

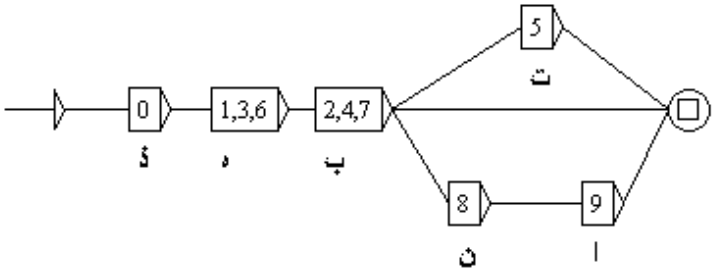
chemins étiquetés par la même chaîne. Comme exemple, nous prenons l'exemple de quelques formes fléchies du verbe arabe « ذهب / aller » qui peuvent être reconnues par l'automate fini A :

Représentation graphique en boites :



Il existe un algorithme de détermination des automates qui transforme tout automate fini non déterministe A en un automate fini déterministe A′ tel que L(A′) = L(A). Soit l'automate A_{dim} fini déterministe issu de l'automate fini non déterministe A de l'exemple précédent :

Représentation graphique en boites :



L'exemple ci-dessus est le résultat de détermination de l'automate précédente. Elle regroupe les états 1, 3, 6 ensembles et les états 2, 4, 7 ensembles.

2.3. Minimisation des automates à nombre fini d'états

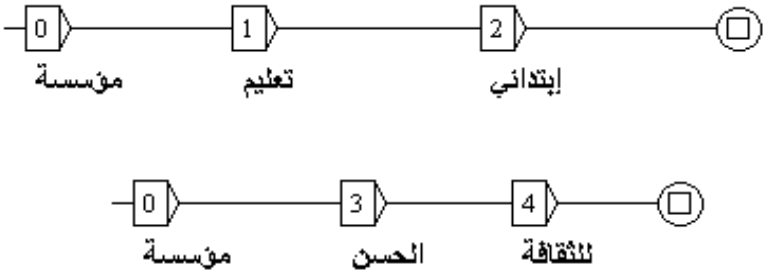
Il existe un algorithme de minimisation pour les automates finis déterministes A afin d'optimiser leur taille et le temps de leur exécution. Nous calculons un automate fini déterministe minimisé A′ tel que L(A) = L(A′) et A′ a un nombre d'états inférieur ou égal au nombre d'état de tout autre automate fini définissant le langage L(A). L'algorithme de minimisation d'un automate fini déterministe A repose sur deux étapes essentielles qui sont la suppression de tous les états inaccessibles d'un automate A et le regroupement de ses états équivalents. Le même exemple illustré précédemment présente un automate fini déterministe minimisé regroupe les états 1, 3, 6 vu qu'ils possèdent la même étiquette. Cet automate contient les états 2, 4, 7 regroupés puisqu'ils ont une étiquette similaire. Nous rappelons que la minimisation d'un automate à nombre fini d'états est une opération qui permet à le transformer en un nouvel automate à nombre fini d'états avec un nombre minimal d'états reconnaissant le même langage donné. Minimiser le nombre d'états est une phase permettant de minimiser la

taille d'un automate réduisant à son tour le temps de réponse lors de la résolution de certains problèmes. L'opération de minimisation s'effectue à travers plusieurs algorithmes consistant à chercher des paires d'états inséparables ou non distinguables comme par exemple l'algorithme de Brzozowski[12].

2.4. Union des automates à états finis

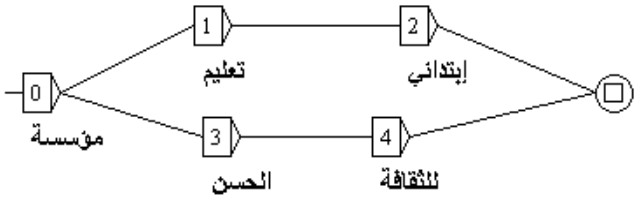
L'union des automates est leur intersection pour unir les états similaires. Cette opération est utilisable pour éviter la redondance au sein des automates construits. Ce genre d'opération est très utilisé dans la reconnaissance des séquences des mots ayant des éléments en commun. Pour illustrer le principe, nous présentons l'exemple suivant :

Représentation graphique en boites :



Après l'union de deux automates à états finis, nous présentons le nouvel automate :

Représentation graphique en boites :



Dans l'exemple précédent, nous avons réuni deux automates qui possèdent le même état 0 étiqueté avec le mot « مؤسسة/ établissement ». Ce mot peut déclencher une ENA ayant deux formes déjà présentées.

2.5. Création d'un conjugueur à la base des automates à états finis

L'automate à états finis est très utilisé pour manipuler les langues peu-dotées à savoir celles européennes comme le macédonien. Dans ce contexte, s'inscrit le travail de [Kostov, 2016] qui consiste à utiliser des algorithmes qui sont des automates à états finis pour élaborer une plateforme appelée FelxiMac1.1[13]. Cette plateforme permet de conjuguer automatiquement des verbes macédoniens dans la plupart des modes et des temps, sans faire appel à une base de

[12] https://fr.wikipedia.org/wiki/Algorithme_de_Brzozowski_de_minimisation_d'un_automate_fini
[13] http://fleximac.free.fr/fra/

données externes. La notion d'automate à nombre fini d'états a été exploitée pour reconnaître le groupe d'appartenance d'un verbe, effectuer une transformation et un réajustement morphologique d'un verbe selon un mode et un temps demandé et à la génération finale de différentes formes verbales.

2.6. Création d'un traducteur à la base des automates à états finis

Le principe des automates à nombre fini d'états plus précisément la transition dynamique entre les états a encouragé les chercheurs à les utiliser dans la traduction automatique pour augmenter la performance des systèmes associés. Dans cette optique, [Semmar et al., 2016] ont pu profiter de ce formalisme dans une partie d'un prototype de moteur de traduction utilisant la recherche d'information interlingue. En fait, les auteurs ont utilisé les automates à nombre fini d'états pour élaborer un formulateur bilingue. Ce formulateur permet de produire pour chaque phrase à traduire un ensemble d'hypothèses de traduction.

3. Réseau de transitions

Les réseaux de transition sont utilisés pour représenter les grammaires locales sous formes des graphes. Ces réseaux sont représentés par des graphes étiquetés dont les nœuds servent à montrer les étapes et les arcs désignant les mots, les catégories lexicales ou syntaxiques. Il existe trois types de réseaux de transition qui sont le réseau de transition simple, le réseau de transition récursif et le réseau de transition augmenté [Lison 2004 ; Kurdi 2018].

3.1. Réseau de transition simple (RTS)

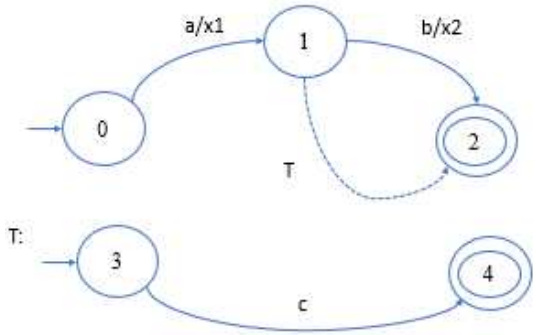
Les RTS, appelés en anglais Simple Transition Networks (STN), rassemblent à un automate à nombre fini d'états. Ce type de réseau de transition peut subir les opérations ensemblistes que nous avons déjà expliquées dans les sous sections (2.2, 2.3, 2.4). L'efficacité des RTS se montre que pour certains types de grammaires. Les RTS ne répondent pas convenablement aux applications du TAL car les règles linguistiques élaborées se caractérisent par leur complexité.

3.2. Réseau de transition récursif (RTR)

Le RTR, appelé en anglais Recursive Transition Network (RTN), est une extension d'un automate à nombre fini d'états [Woods 1970]. Un RTR a la forme d'un graphe composé des nœuds et des transitions. Chaque transition dans ce graphe peut atteindre un nœud final ou non-final. Le RTR peut traiter un nœud non-final comme un appel à un autre RTR ce qui fait sa différence par rapport à un automate à nombre fini d'états. Formellement, un RTR est défini par un sextuple $(Q, I, \Sigma, q_0, F, \delta)$, où :

- Q est un ensemble fini et non vide d'états,
- I est l'ensemble des états sous-initiaux (états qui étiquettent au moins une transition du transducteur RTN, et représentent un appel récursif au sous-RTN),
- Σ est un alphabet de symboles complexes. Chaque symbole est constitué d'une paire (a,b) dont a appartient à l'alphabet d'entrée E et b à celui de sortie S,
- q_0 est un état initial appartenant à Q,
- F est un ensemble non vide de Q appelé l'ensemble des états finaux,
- $\delta : Q \times (\Sigma \cup I \cup \{\varepsilon\}) \rightarrow Q$ est une fonction de transitions.

Nous proposons le RTR suivant dont la transition T fait appel à un sous RTR. Le RTR principal est défini par Q : {0,1,2,3,4}, I : {0,3}, q_0 :{0} qui est l'état initial du graphe principal, F :{2,4} et Σ : {(a, x1), (b, x2), (c, ε)}.



Les RTR peuvent être utilisés pour extraire les EN. Pour la langue arabe, nous proposons un chemin d'extraction d'une EN ayant la catégorie *évènement culturel*.

Représentation graphique en boîtes :

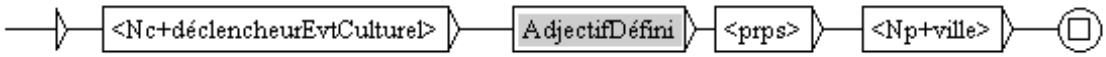
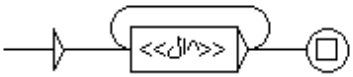


Illustration de la boîte grise :



Le chemin d'extraction d'une EN arabe désignant un évènement culturel part d'un nœud initial et rencontre la première boîte contenant un trait grammatical (se trouvant dans un dictionnaire). Il passe par la suite à la boîte grise faisant l'appel à un sous-graphe appelé « Adjectif Défini ». Le parcours du chemin d'extraction principale va continuer après avoir lu les boîtes du sous-graphe déjà mentionné jusqu'à atteindre le nœud final du graphe principal. L'EN arabe « مهرجان

الطرب الأندلسي بالمغرب » peut être extraite via le graphe principal illustré sachant que les deux adjectifs définis « الطرب الأندلسي » ont été reconnue grâce au sous-graphe.

3.3. Réseau de transition augmenté (RTA)

Les RTA, appelés en anglais Augmented Transition Network (ATN), sont des RTR étendus [Woods 1970]. Les nouvelles extensions aux RTA constituent à ajouter une structure de données appelée Registre permettant de conserver les informations, à obliger aussi des conditions sur les transitions et à associer des actions aux transitions effectuées [Bates 1978]. Les actions associées peuvent être utiliser pour modifier la structure des informations à renvoyer en sortie.

Représentation graphique en boîtes :

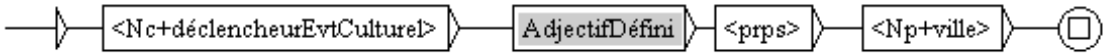


Illustration de la boîte grise avec une condition :

Dans l'exemple illustré ci-dessus, la condition est ajoutée pour augmenter les critères de sélection du trait grammatical à sélectionner dans le sous graphe en boîtes grises. Dans ce cas, la condition a la forme d'un intervalle de reconnaissance ; au minimum zéro adjectif défini et au maximum deux adjectifs définis.

4. Transducteurs à nombre fini d'états

Les transducteurs à nombre fini d'états permettent de traiter des phénomènes complexes de la langue à savoir la flexion, la désambiguïsation et la résolution de l'agglutination. Dans ce qui suit, nous donnons leur définition formelle ainsi que leur représentation graphique (diagramme de transitions et en boîtes).

4.1. Définition formelle d'un transducteur

Un transducteur à nombre fini d'états à une entrée alphabet Σ_i et une sortie alphabet Σ_o est un sextuple $(Q, \Sigma_i, \Sigma_o, q_0, F, \delta)$ tel que :

- Q est un ensemble fini et non vide des états,
- Σ_i est l'alphabet d'entrée (ensemble fini et non vide),
- Σ_o est l'alphabet de sortie (ensemble fini et non vide),

- q_0 est un état initial appartenant à Q,
- F est un ensemble de Q appelé l'ensemble des états finaux,
- δ est une fonction de transitions qui associe un état initial q_1 appartenant à Q et un mot d'entrée w_i de Σ_i^* avec un état d'arrivé q_2 de Q et un mot de sortie $\sum_o^*$ et on note δ (q_1, w_i, q_2, w_o).

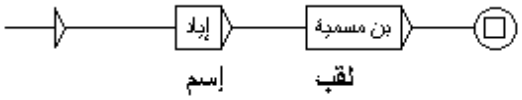
Un transducteur à nombre fini d'états est un automate à nombre fini d'états dont les transitions sont étiquetées avec un couple de symboles, un symbole reconnu en entrée et un symbole produit en sortie. Autrement dit, un transducteur permet de reconnaître une chaîne d'entrée et produit en sortie, un jeu de caractères différent.

4.2. Représentation graphique d'un transducteur

L'adaptation d'une représentation graphique pour les transducteurs assure une meilleure lisibilité aux grammaires locales décrites sous une forme formelle. Chaque transducteur est caractérisé par un nœud initial et un nœud final. Il est composé aussi des états et des transitions sachant qu'un état peut être un nœud simple et vide.

Diagramme de transitions :

Représentation graphique en boîtes :



Le transducteur illustré permet de reconnaître en entrée deux chaines de caractères et de générer en sortie deux nouvelles chaînes. Ce format graphique se réalise sous des outils ou des plateformes linguistiques spécifiques.

4.3. Traduction automatique à la base des transducteurs

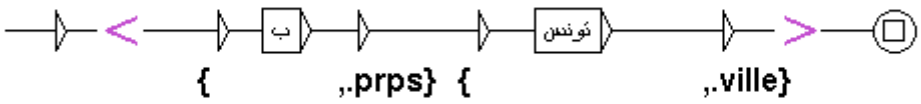
Les transducteurs peuvent être exploités pour réaliser la tâche de traduction automatique des EN d'une langue à une autre. En fait, cette tâche facilite également l'enrichissement du volume manquant dans les lexiques bilingues des EN.

Représentation graphique en boites :

La traduction réalisée par ce transducteur consiste à remplacer l'ancienne séquence de mots « كلية العلوم بتونس » par son équivalent en français qui est « Faculté des Sciences de Tunis ».

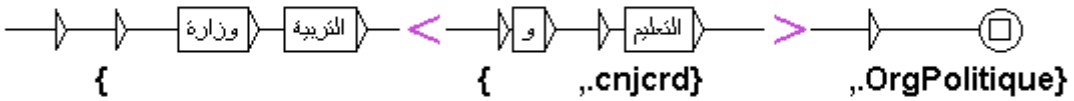
4.4. Résolution de l'agglutination à la base des transducteurs

Les transducteurs sont capables de traiter l'agglutination selon les fonctionnalités avancées de la plateforme exploitée. L'agglutination peut être une préposition ou encore une conjonction attachée à un mot. Dans notre cas, Unitex traite l'agglutination selon le mode morphologique qui exploite les indicateurs < et >.



D'après l'exemple illustré, nous constatons que le transducteur a reconnu en mode morphologique la préposition « ب » qui est attachée à un nom de ville.

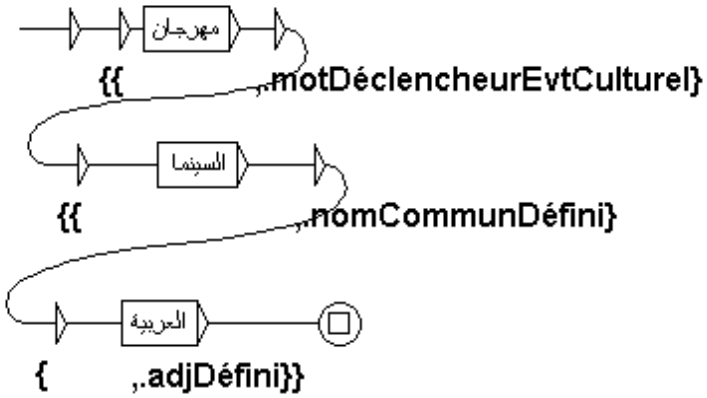
Dans la langue arabe, la conjonction est toujours attachée au mot qui la précède. Dans ce cas, les transducteurs sont très fiables puisqu'ils permettent de lire cette conjonction.



D'après cet exemple, le transducteur lit la conjonction de coordination et il la protège par une balise d'annotation. Ce transducteur ne peut pas agir sur cette conjonction avec une simple lecture mais avec une lecture morphologique.

4.5. Reconnaissance des syntagmes dans des EN via des transducteurs

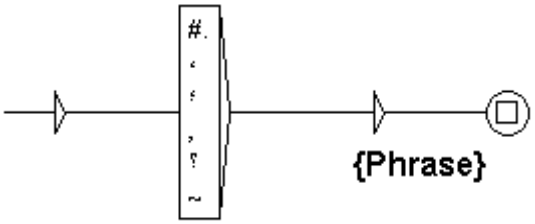
Les transducteurs permettent l'analyse syntagmatique lors de la reconnaissance des EN. Dans ce contexte, nous présentons un transducteur qui reconnaît et annote un syntagme adjectival.



Le transducteur illustré reconnaît la séquence de mot qui compose une ENA décrivant un nom d'évènement culturel. Après le mot déclencheur, le transducteur reconnaît un nom commun et il l'annote selon sa catégorie et sa définition. Puis, il reconnaît un adjectif qui est aussi défini dépendant du nom commun déjà détecté.

4.6. Segmentation des textes à la base des transducteurs

Les transducteurs servent également à la segmentation des textes selon la langue traitée. La segmentation varie selon les exigences des utilisateurs. Pour cette raison, nous trouvons que les balises d'annotations diffèrent selon des préférences personnelles.



Le transducteur illustré permet de segmenter un texte arabe selon les signes de ponctuations présentées dans le premier nœud. La segmentation dans ce cas vise à délimiter le texte pris en entrée par le transducteur en des phrases.

Après leur établissement, les transducteurs nécessitent un ordre de passage pour qu'ils agissent sur le texte avec une bonne précision et un bon rappel car plusieurs systèmes basés sur ce formalisme souffrent d'avoir une bonne précision mais un mauvais rappel. De même, le passage d'un ensemble de transducteurs arbitrairement peut influencer sur le coût en termes de temps d'exécution.

5. Cascade de transducteurs

Le principe d'une cascade de transducteurs est défini comme une succession de transducteurs appliqués à un texte dans un ordre spécifique afin de convertir ou extraire des motifs. Chaque transducteur se base sur les résultats de ses prédécesseurs dans la même cascade. L'ordre de passage des transducteurs dépend de leur degré de certitude [Abney, 1996 ; Friburger, 2002].

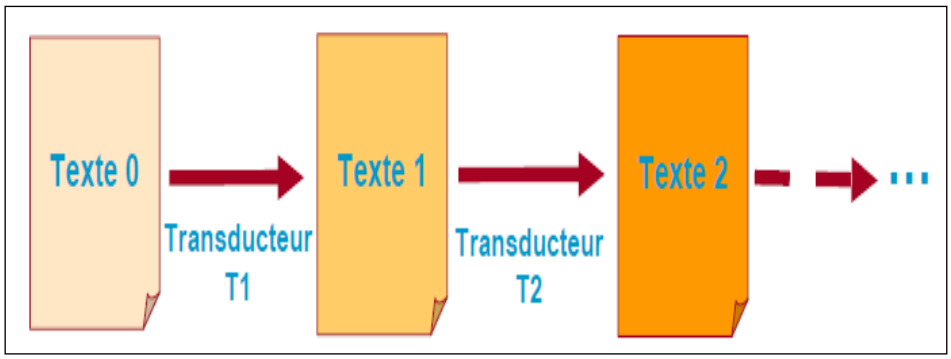


Figure 3. Principe d'une cascade des transducteurs

La figure 3 décrit le principe d'une cascade de transducteurs schématiquement en partant d'un texte original (brut) jusqu'à avoir un texte modifié par le transducteur final. Nous constatons que chaque transducteur faisant partie de la cascade agit, dans l'instant t_i, sur le texte noté T_i résultant de l'application de son prédécesseur dans l'instant t_{i-1}.

Le pouvoir d'une cascade de transducteurs est de réutiliser des motifs déjà reconnus ou d'éviter leur chemin de reconnaissance. De plus, elle assure un temps de traitement respectable au fait qu'il est compatible avec des applications de vraie grandeur. Au sein de la cascade, la succession des transducteurs définie fournit des qualités intéressantes en termes de réutilisabilité et de modularité. Le regroupement de transducteurs au sein d'une cascade n'est pas un choix aléatoire. Il a comme objectif d'ordonner leur passage d'application sur le texte. En outre, nous commençons par l'application d'un transducteur permettant de trouver les chemins les plus sûrs. Ces chemins reconnaissent les motifs les plus évidents d'une part et les moins ambigus d'une autre part. En fait, l'ordre du passage permet d'ajouter non seulement un degré de certitude mais également une réduction du champ de recherche pour le reste des transducteurs regroupés dans la même cascade. De plus, il permet d'optimiser la reconnaissance d'une certaine séquence en diminuant le coût du temps d'exécution.

6. Systèmes basés sur les cascades de transducteurs

Il existe plusieurs systèmes basés sur les cascades de transducteurs développés pour aborder les tâches suivantes : l'analyse syntaxique, l'extraction de l'information et de la traduction automatique. Ces systèmes ont été réalisés en tirant profit des avantages du principe d'une

cascade de transducteurs en termes de robustesse, précision de résultats et rapidité d'exécution. Dans ce qui suit, nous rappelons quelques systèmes.

6.1. Cascades de transducteurs pour l'Extraction d'Information

Pour l'EI, un système appelé FASTUS a été développé par [Hobbs et al., 1993]. Ce système traite des textes en anglais et en japonais. Dans son architecture, FASTUS se base sur le principe d'une cascade de transducteurs à nombre fini d'états afin d'extraire des informations pertinentes. Dans la première étape à effectuer par ce système concerne le traitement lexical permettant de reconnaître des mots complexes tels que les expressions multi-mots et des noms propres. Ensuite, après avoir effectué l'analyse morphologique, le système passe à analyser syntaxiquement les groupes nominaux simples et complexes et de la même façon pour les groupes verbaux. Après, le système traite le niveau sémantique en reconnaissant les évènements du domaine à travers des patrons d'extraction construits. En somme, le système fusionne les informations extraites à partir des textes si elles concernent la même entité.

Dans le cadre de FACILE [Ciravegna et al., 1999], un projet financé par l'Union européenne traitant la classification des textes et l'extraction d'informations en domaine financier, [Ciravegna et Lavelli, 1999] ont mis en œuvre une cascade de transducteurs pour l'extraction des informations en trois cascades successives. Ces cascades contiennent respectivement des règles empiriques, des cas réguliers de la grammaire et des règles applicables si seulement si aucune règle de deux cascades précédentes n'a travaillé.

Pour la REN, Les auteurs de [Maurel et al., 2011] ont proposé un système basé sur une cascade de transducteur, appelé CasEN. Ce dernier est implanté à travers l'outil CasSys intégré sous la plateforme linguistique libre Unitex. CasEN repose sur un ensemble de transducteurs agissant sur le texte par des insertions, remplacements ou suppressions. Ces transducteurs sont répartis en cinq catégories de graphes : les graphes de reconnaissances, les graphes d'outils, les graphes de listes, les graphes de masques et les graphes étiqueteurs.

Dans le cadre du projet Biosystémique, [Landomiel at al., 2017] ont exploité Unitex pour proposer une cascade subdivisée en trois sous-cascades pour découvrir des relations entre protéines. Pour créer les graphes déjà mentionnés, les auteurs ont pu profiter de plusieurs options offertes par Unitex telles que le mode morphologique et le contexte négatif.

6.2. Cascades de transducteurs pour la fouille de texte

Un outil appelé LIZARD a été développé par [Balvet, 2002]. C'est un outil qui vise à assister les développeurs de ressources linguistiques en automatisant la fouille de corpus. Il est basé sur les analyses partielles et les cascades de transducteurs à nombre fini d'états. Par le biais de la

plateforme Intex [Silberztein et al., 2001], la cascade de transducteurs permet de prétraiter, étiqueter puis explorer des corpus de textes de façon efficace.

6.3. Cascade de transducteurs pour la segmentation de parole

Dans le cadre d'un projet appelé ANR EPAC, [Mokrane et al., 2008] ont développé un système pour la segmentation de la parole conversationnelle. Ce système permet d'indexer et annoter automatiquement des grands flux de paroles issues d'émissions télévisées ou radiophoniques. La segmentation est basée sur une cascade de transducteurs qui identifie dans une première passe les segments (chunks) ayant une structure normalisée qui suivent une stratégie par ilots de confiance. La seconde cascade de transducteurs se limite à la caractérisation des catégories complémentaires aux annotations PEAS ; le chunk COO et le chunk PONCT (pour la ponctuation). Elle attribue enfin l'étiquette CHINC (chunk inconnu) aux zones non encore segmentées. Ces séquences seront par la suite analysées, soit pour caractériser les diffluences (ce sont les zones non segmentées), pour corriger les erreurs de reconnaissance et les erreurs d'étiquetage morphosyntaxique.

7. Unitex

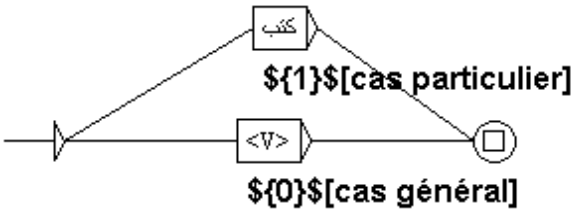
Unitex[14] est une plateforme linguistique libre développée par Sébastien Paumier en 2002. En fait, Unitex consiste à traiter des ressources textuelles en des langues naturelles via un environnement de travail manipulable graphiquement ou à travers des lignes de commande [Paumier, 2017]. De plus, le traitement d'une ressource textuelle s'effectue à travers des ressources lexicales et des grammaires locales. Unitex est connu par son multilinguisme car elle regroupe plusieurs langues apparaissant sous formes de dossiers lors de son téléchargement. Unitex offre l'opportunité de manipuler des textes bruts qui ne subissent aucun processus de prétraitement. Les textes candidats n'admettent pas une contrainte spéciale pour cerner la limite de leur taille. En outre, quel que soit la taille d'un texte donné, Unitex permet de le charger en proposant une phase de prétraitement si nécessaire à travers une fenêtre. Dans cette dernière, il existe quelques options proposées pour orienter l'utilisateur vers le type du prétraitement adéquat. Il est possible de charger un texte balisé aussi en choisissant une option particulière dans la fenêtre déjà mentionnée pour ne pas perturber l'application des prétraitements.

7.1. Transducteurs sous Unitex

Étant donné qu'une grammaire locale est un moyen puissant de représenter quelques phénomènes linguistiques, Unitex intègre la notion de transduction empruntée aux automates à nombre fini d'états. Les transducteurs sous Unitex peuvent être schématisés via plusieurs types

[14] http://unitexgramlab.org/

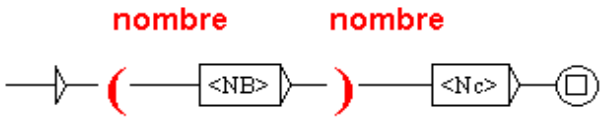
de boîtes. A leur tour, les boîtes ne sont pas dédiées seulement à stocker les traits de dictionnaires mais également à produire des sorties. Il est possible d'attribuer un poids aux boîtes d'un transducteur pour choisir un chemin adéquat ayant un poids maximal lors qu'il existe des sorties différentes associées plusieurs chemins pour la même séquence reconnue.



Il faut mentionner que la notion de poids n'est valable qu'à l'intérieur du graphe. Toutefois, elle n'est pas valide dans les sous-graphes et les graphes appelants.

7.1.1. Notion de variables dans les transducteurs

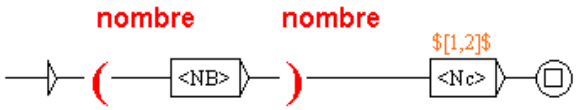
Les sorties sur les chemins d'un transducteur peuvent être organisées à travers la notion des variables. En fait, les variables d'entrées permettent de sélectionner juste des parties souhaitées du texte reconnu par ce transducteur.



L'appel d'une variable définie dans le graphe comme « nombre » se fait en encadrant son nom avec le caractère $.

7.1.2. Notion de répétition dans les transducteurs

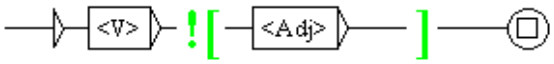
Une boîte dans un transducteur peut avoir un intervalle fixant son nombre de répétition. Par exemple, si nous fixons un intervalle [m, M] à une boîte (nom commun), le chemin reconnaît des séquences avec au moins m noms commun consécutifs et pas plus de M noms commun.



Après la lecture d'un nombre dans le graphe illustré dans l'exemple, nous rencontrons une séquence qui doit contenir au moins nom commun et au maximum deux noms communs.

7.1.3. Notion de contexte négatif

Parmi les utilisations avancées d'un transducteur, nous pouvons citer la notion d'un contexte négatif. Cette notion permet de fixer une condition d'arrêt en analysant le contexte d'une séquence à reconnaître.



L'exemple ci-dessus décrit un chemin reconnaissant un verbe non suivi par un adjectif.

7.1.4. Notion de filtre et mode morphologique

Un transducteur sous Unitex peut fonctionner avec des boîtes utilisant des filtres morphologiques pour effectuer des requêtes entrant à l'intérieur d'un token. Cependant, l'utilisation de ce genre de filtre ne permet pas d'exploiter les traits de dictionnaires. Pour cette raison, Unitex offre un mode morphologique avec lequel les boîtes d'un transducteur peuvent découper un mot et faire référence aux traits d'un dictionnaire. Ce mode s'exprime à travers les indicateurs < et >. Comme elle indique la figure suivante, le mode morphologique exige le choix des dictionnaires à exploiter avec l'extension « .bin » dans le menu « Info » et la fenêtre « Morphological-mode dictionaries ».

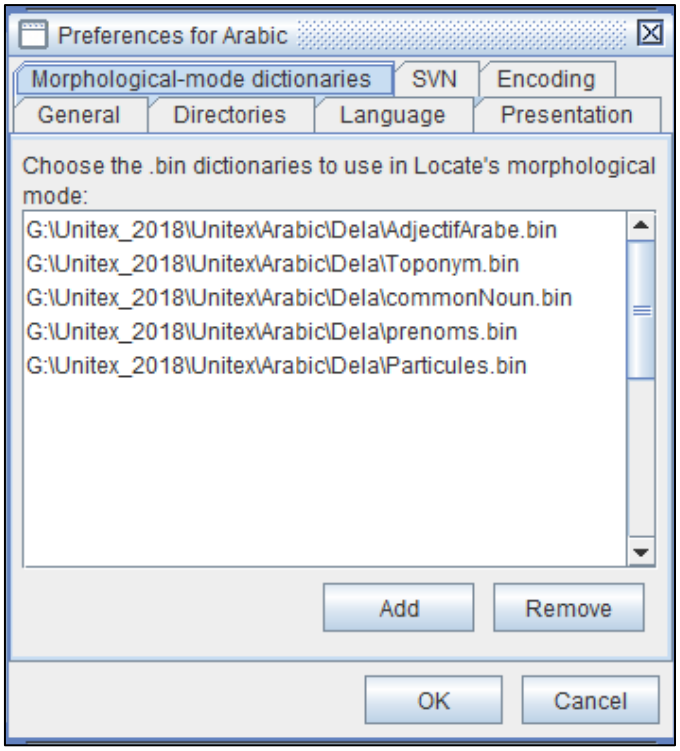


Figure 4. Ajout des dictionnaires pour le mode morphologique

7.2. Création d'une cascade de transducteurs sous Unitex

Sous Unitex, la création d'une cascade de transducteurs constitue à appeler plusieurs transducteurs (graphes) avec un ordre précis pour les appliquer sur un texte. Chaque transducteur appelé effectue des modifications sur ce texte selon des objectifs fixés. Unitex offre un outil intitulé CasSys dédié à créer une cascade de transducteurs.

7.2.1. CasSys

Le premier modèle de l'outil CasSys a été créé en 2002 au laboratoire LI (Laboratoire d'informatique de l'université de Tours) [Maurel et al., 2013]. Cet outil était initialement dédié à la REN et il était généralisé pour effecteur d'autres traitement.

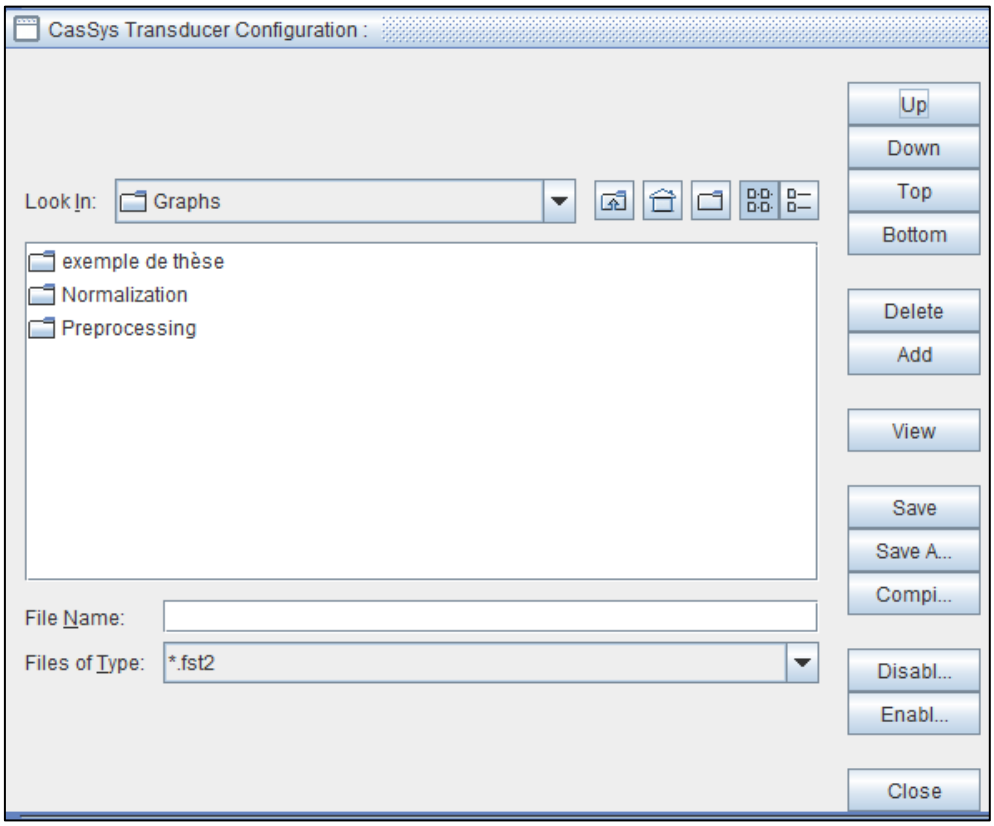


Figure 5. Interface de l'outil CasSys

D'après la figure 5, l'interface de CasSys permet de configurer la cascade à créer selon les graphes déjà conçus. Encore, nous trouvons des boutons spécifiques à manipuler l'ordre des graphes comme « Up » et « Down ». La sauvegarde et la compilation sont fournies aussi sous formes de boutons. Nous pouvons visualiser les graphes regroupés au sein de la cascade à générer grâce au bouton « View ». Une cascade de transducteurs applique plusieurs graphes compilés l'un après l'autre sur le texte. En outre, chaque graphe va modifier le texte selon son rôle. De plus, les changements effectués sur un texte à l'instant t=0 peuvent être utilisés pour des traitements supplémentaires par les graphes suivants.

7.2.2. Application et modes de passage d'une cascade de transducteurs

Le choix application d'une cascade de transducteurs crée via l'outil CasSys n'est pas situé sur son interface. Pour l'appliquer, il faut aller au menu « texte » et choisir « Apply CasSys Cascade » comme elle indique la figure suivante.

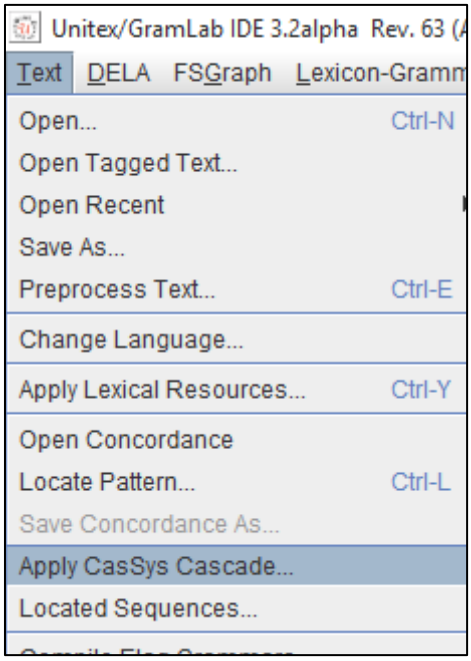


Figure 6. Application d'une cascade de transducteurs

Lors de l'application d'une cascade de transducteurs, il est possible de choisir le mode de comportement de chaque transducteur sélectionné (Figure 7). Chaque mode choisi a des particularités et certains modes ne peuvent pas être utilisés en parallèle avec un autre mode.

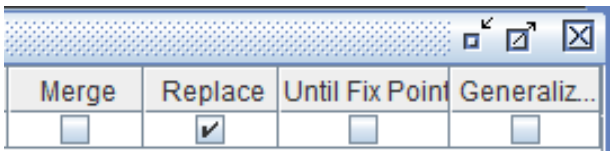


Figure 7. Modes de passage d'une cascade de transducteurs

Les deux modes ordinaires de CasSys sont « merge » et « replace » dont le premier fusionne l'entrée avec la sorte qui sera insérée à gauche de la séquence reconnue. Le mode « replace » ne peut pas être choisi pour un transducteur au même temps que le premier mode car celui-ci permet de remplacer la séquence reconnue avec la sortie.

Un nouveau mode est utilisé par CasSys qui inspire de la structure itérative dont le nombre d'itérations est inconnu. Ce mode est appelé « Until fix point » permettant d'appliquer un transducteur sur un texte de manière itérative tant que de nouvelles concordances peuvent être obtenues. Autrement dit, l'application d'un transducteur s'arrête lorsque le texte traité n'admet aucun changement.

Le dernier mode pouvant être choisi dans CasSys s'appelle « Generic » qui est lié à un transducteur de généralisation d'étiquetage dont il exploite les séquences reconnues dépendant d'un contexte précis pour récupérer les mêmes séquences apparaissant hors de leur contexte déjà mentionné.

7.3. Comparaison entre Unitex et NooJ

NooJ[15] est un environnement de développement linguistique créé par Max Silberztein en 2002 [Silberztein, 2002]. Cet environnement intègre plusieurs outils de traitement automatique offrant la possibilité de traiter les corpus ayant plusieurs formats comme les fichiers html et PDF. Comme Unitex, NooJ permet de construire ou mettre à jour les dictionnaires (en traitant la morphologie, la flexion, etc.) et d'explorer les corpus [Hammouda et Haddar, 2017 ; Hammouda et Haddar, 2016]. Le choix d'utilisation de l'une des deux plateformes sur de différents critères. Pour cette raison, nous proposons le tableau suivant.

Tableau 5. Comparaison entre Unitex et NooJ

	Unitex	**NooJ**
Accessibilité	Gratuit	Gratuit
Licence	LGPL	LGPL
Dernière version	Unitex 3.2 alfa	NooJ v5.0
Utilisation	Académique et commerciale	Commerciale
Type	Analyse et annotation	Annotation
Création des grammaires	Règles et graphes	Règles et graphes
Multilinguisme	22 langues	23 langues
Format de corpus d'étude	Txt (Unicode)	Plus de 100 formats
Format de corpus de sortie	Html, XML, Csv et txt	Html, XML et Csv
Création de noyau	C++ (portabilité)	C++, Java
Création d'interface	Anglais/ Java	Anglais/ Java
Analyse morphologique	Filtre et mode	Mode
Technologie de graphes	Généralisation d'étiquetage Notion de contexte Notion de répétition des boîtes	-
Mode-passage de graphes	4 modes	-
Génération d'une cascade	Outil CasSys avec interface conviviale	Module intégré

[15] www.nooj-association.org (consulté le 27/04/2018)

Le tableau 5 montre que les deux plateformes possèdent des points en commun comme la licence libre. NooJ propose des fonctionnalités au niveau des extensions des textes à manipuler mais elle souffre au niveau des techniques avancées exploitées pour traiter cette variété des textes. La différence réside aussi au niveau des graphes qui sont caractérisés par leur progression sous Unitex pour qu'ils s'appliquent avec divers modes de passage assurés par l'outil de génération des cascades. L'obtention d'une cascade de transducteurs sous NooJ s'agit d'une simple liste de la version compilée des graphes sans avoir un mode particulier. Cette option se trouve dans un menu et n'admet aucune indication pour y pointer directement.

8. Conclusion

Dans le présent chapitre, nous avons rappelé la définition formelle d'un automate à nombre fini d'états et son principe de fonctionnement. Parmi les types d'automates finis, nous avons abordé les automates à nombre fini d'états en commençant par sa représentation formelle et son déterminisme. De plus, nous avons présenté les opérations sur ce type d'automates pour en finir avec les applications qui exploite son principe. Après, nous avons expliqué les trois types qu'un réseau de transitions peut admettre. En fait, nous avons focalisé sur le réseau de transitions augmentés vu qu'il se représente sous forme d'un graphe admettant des conditions spécifiques. L'évolution d'un automate a mené à la notion de transducteur qui admet également une représentation formelle qui le caractérise. Nous avons montré l'importance des transducteurs en illustrant diverses applications. Puis nous sommes passés à la notion de cascade de transducteurs et son principe. Les applications basées sur les cascades de transducteurs sont nombreuses de sorte qu'elles appartiennent à divers domaines comme l'EI. Nous avons clôturé ce chapitre par une section dédiée à la représentation de plateforme linguistique Unitex en effectuant une courte comparaison avec la plateforme NooJ.

Partie 2 : Etude linguistique

Chapitre 3 : Typologie des entités nommées arabes

L'étude linguistique pour définir une typologie d'ENA vise chercher les termes et les expressions pertinentes faisant référence à une EN dans un corpus extrait de la Wikipédia arabe. La recherche des ENA que nous voulons effectuer permet d'avoir une définition d'EN précise et concrète pour faciliter leur reconnaissance. Dans notre cas, nous tentons de dégager les ENA et de les catégoriser afin de chercher ultérieurement les relations qui les relient. Pour cette raison, nous devons profiter des ENA délimitées pour la détermination de leur catégorie d'appartenance, ce qui favorise l'élaboration d'une typologie de catégories. Notamment, la typologie peut subir un raffinement pour qu'elle se compose encore de sous-catégories afin augmenter le niveau de granularité. Durant notre étude linguistique pour d'identifier les ENA, nous essayons de détecter plusieurs formes d'ENA d'une part et de prédire les autres formes selon leur contexte d'apparition d'une autre part.

Le chapitre courant se compose de trois sections principales dont la première présente l'identification des ENA à effectuer en se basant sur un corpus d'étude extrait à partir de la Wikipédia arabe. Dans cette phase de recherche, nous commençons par la définition d'ENA retenue pour délimiter ses éléments. Ensuite, nous décrivons la hiérarchie d'ENA que nous créons englobant les catégories et les sous-catégories identifiées. Dans la deuxième section, nous présentons les processus de catégorisation effectuée pour dégager les catégories et les sous-catégories d'appartenance d'une ENA. Pour chaque catégorie identifiée, nous présentons les formes associées à une ENA illustrées par des exemples. Dans la troisième section, nous étudions les formes imbriquées d'une ENA et l'appel de plusieurs catégories et sous-catégories au sein d'une même ENA. Finalement, nous clôturons par une conclusion.

1. Identification des ENA

La recherche les ENA signifie l'analyse de toutes ses formes pouvant apparaitre dans le corpus d'étude extrait à partir de la Wikipédia arabe. Les formes alternatives d'une ENA possèdent différentes écritures régionales ce qui favorise leur détection. A travers cette recherche, nous visons à explorer notre corpus d'étude pour diverses raisons. La première raison est l'analyse des expressions décrivant une forme d'ENA afin de la délimiter. La deuxième raison consiste à catégoriser chaque ENA délimitée pour deviner sa classe d'appartenance. Réellement, la délimitation et la catégorisation des ENA détectées ne sont pas des tâches faciles. En fait, ces deux tâches nécessitent l'étude du contexte d'apparition de chaque ENA ainsi les mots déclencheurs associées qui jouent le rôle des indices de repérage. La délimitation d'une ENA exige la présence d'une définition précise pour cerner ses éléments. Pour cette raison, nous présentons la définition retenue pour détecter une ENA et faciliter sa catégorisation.

1.1. Définition d'ENA retenue

Avoir une définition d'une ENA joue un rôle très important pour bien délimiter ses composantes. En fait, une définition pertinente peut traiter toutes les formes alternatives des ENA selon les différentes catégories d'appartenance. Autrement dit, la définition d'une ENA constitue l'abstraction de ses formes. Dans ce contexte, nous proposons la définition suivante pour explorer et analyser les ENA apparaissant dans notre corpus d'étude :

> *« Une ENA est une expression qui peut ne pas contenir un nom propre et elle structurée à travers des catégories et des sous-catégories. »*

Par conséquent, nous traitons l'ENA non seulement si elle est une expression qui doit contenir forcément un nom propre mais nous la considérons également en tant qu'une expression qui peut contenir un ensemble de constituants ayant différent nature (nom commun, adjectif, etc.). De plus, nous pouvons structurer une ENA à travers les catégories ayant à leur tour un ensemble de sous-catégories. Après la proposition d'une définition d'EN, nous nous trouvons face à la catégorisation qui est une étape délicate dépendant de plusieurs opinions. Dans ce qui suit, nous décrivons la hiérarchie d'ENA établie.

1.2. Hiérarchie d'ENA établie

L'étude de différentes formes d'ENA nous a aidé à effectuer une catégorisation profonde. Nous rappelons que cette catégorisation n'est pas faite aléatoirement mais elle s'est basée sur une définition claire et précise d'une ENA. Par conséquent, nous avons élaboré une hiérarchie représentant schématiquement la catégorisation réalisée.

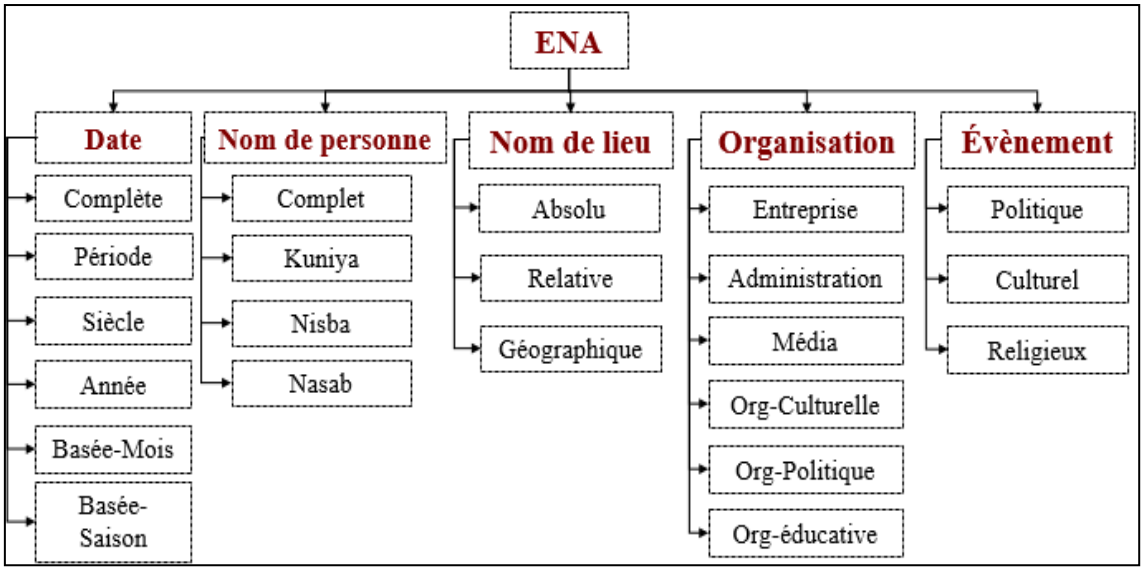


Figure 8. Hiérarchie d'ENA établie

La figure 8 décrit notre hiérarchie d'ENA élaborée après l'exploration et l'analyse de notre corpus d'étude. Les catégories principales sont raffinées pour qu'elles soient décomposées en sous-catégories. Nous avons étendu les catégories proposées pour offrir un niveau de granularité facilitant par la suite l'étude de la liaison sémantique entre ces ENA. Cette extension dépend en large partie de différentes formes d'ENA dans le corpus d'étude.

En fait, notre contribution se focalise non seulement sur le fait de dégager des catégories simples, mais également sur une sous-catégorisation raffinée pouvant avoir à leur tour d'autres sous-catégories. Dans ce cas, nous avons touché trois niveaux de raffinement. Prenons l'exemple de la catégorie nom de lieu, celle-ci possède trois sous-catégories parmi lesquelles nous citons le nom de lieu relatif ayant 16 autres sous-catégories descendantes. Etant donné que la catégorie nom de lieu possède trois sous-catégories qui sont raffinées, alors chaque sous-catégorie fait appels à trois sous-hiérarchies. Dans ce qui suit, nous décrivons celle dédiée au nom de lieu absolu avec des quelques exemples illustratifs.

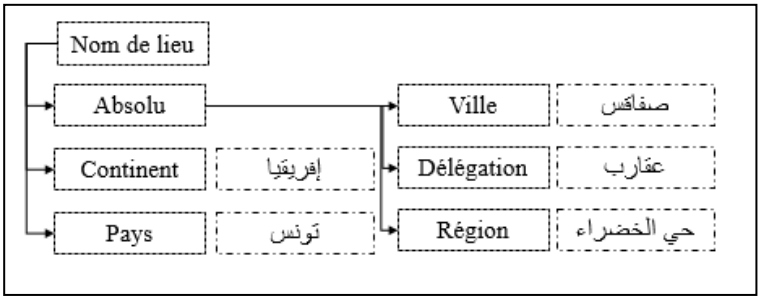


Figure 9. Sous-hiérarchie décrivant les sous-catégories d'un nom de lieu absolu

La figure 9 décrit quelques instances associées aux sous-catégories d'un nom de lieu relatif. Ces instances peuvent être ou ne pas être équipées par des mots déclencheurs. Parfois, nous les détectons agglutinés d'où vient la nécessité de les séparer. D'ailleurs, l'agglutination touche les mots déclencheurs aussi. Décrivons maintenant les deux autres sous-hiérarchies.

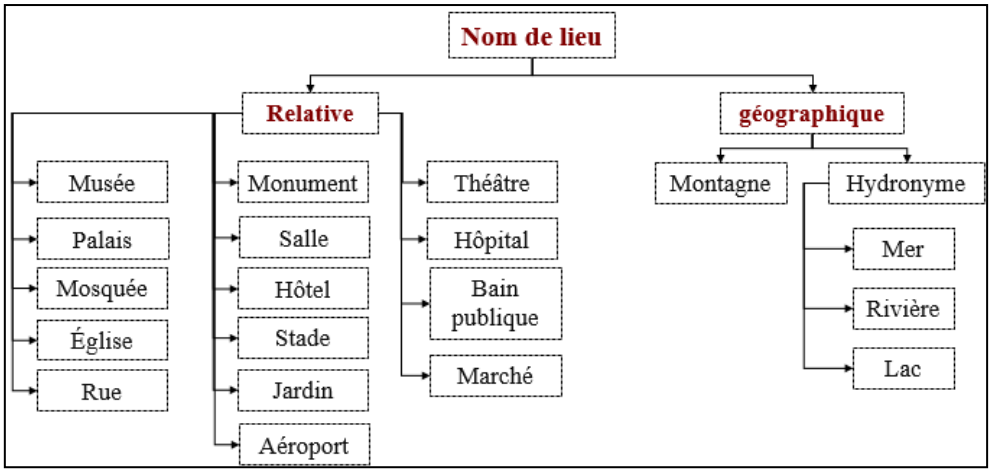


Figure 10. Sous-hiérarchie décrivant les sous-catégories : relative et géographique

La figure 10 illustre les deux sous-typologies dont la première est dédiée au nom de lieu relatif. Cette dernière regroupe 16 sous-catégories ayant différents chemins de détection d'une ENA et des mots déclencheurs variés. Ces sous-catégories peuvent recevoir même les anciennes ENA vu qu'il existe des textes historiques dans notre corpus d'étude. Notamment, la sous-catégorie nom de lieu géographique possède deux branches, soit la sous-catégorie Montagne, soit la sous-catégorie Hydronyme. Nous constatons que le raffinement touche un niveau égal à 4 puisque Hydronyme englobe 3 sous-catégories qui sont Mer, Rivière et Lac.

Dans la section suivante, nous allons décrire les catégories appartenant à la hiérarchie principale retenue. Pour chaque catégorie, nous illustrons les sous-catégories déjà mentionnées à travers des exemples.

2. Catégorisation d'ENA

La catégorisation est une étape visant déterminer la catégorie adéquate qui décrit convenablement une ENA. Pour deviner les catégories d'appartenance, nous nous basons sur les mots déclencheurs qui précèdent ou qui suivent une ENA délimitée. De plus, nous exploitons également ces mots déclencheurs pour dégager les sous-catégories. En cas d'absence des mots déclencheurs, nous analysons le contexte d'apparition de l'ENA à repérer. Notre catégorisation donne naissance à cinq catégories principales : Date, Nom de personne, Nom de lieu, Evénement et Organisation. Dans ce qui suit, nous présentons ces catégories ainsi les sous-catégories qu'elles possèdent avec des exemples explicatifs.

2.1. Catégorie Date

La catégorie Date qui décrit une ENA fait partie des expressions numériques. Dans notre corpus d'étude, nous avons trouvé que les formes suivantes pouvaient décrire une ENA date : période, siècle, année, date basée sur le mois, une date complète ou une saison suivie par une année. Dans ce qui suit, nous décrivons les formes mentionnées avec des exemples illustratifs.

La période est une forme parmi celles décrivant une date. Elle peut être calculée en se basant sur plusieurs opérandes comme le mois, le jour, le siècle ou l'année. Cette forme peut être identifiée soit via les indicateurs (يومي / les deux jours) soit via une préposition décrivant la notion d'un intervalle de temps par exemple « من /de ». L'étude que nous avons faite montre que certains indicateurs peuvent apparaitre en différentes formes morphologiques (pluriel, duel, etc.) comme le mot « سنتي / les deux années » en (3) or une conjonction d'années précédées par le mot « الأعوام /les années ».

(1) من 17 إلى 19 ديسمبر

De 17 au 19 décembre

(2) من القرن 11 م إلى القرن 17 م

De 11ème au 17ème siècle

(3) سنتي 1999 و2000

Entre les deux années 1999 et 2000

Un siècle est un cycle d'années utilisé pour décrire une longue période. Généralement, Cette forme apparait dans les articles ayant une nature historique dans notre corpus d'étude. Une ENA exprimant un siècle figure toujours avec l'indicateur « القرن/ le siècle » faisant partie d'elle comme elles sont illustrées en (4) et (5). L'indicateur déjà mentionnée est peut-être associé à une préposition jouant le rôle d'un indicateur de temps. En fait, la présence d'une préposition permet de réduire la complexité de sa détection. Il faut mentionner aussi que le nom de siècle peut être écrit sous forme de chiffres ou en toute lettre.

(4) في القرن 11 م

Au 11ème siècle

(5) في القرن التاسع للميلاد

Au 9ème siècle

Parmi les formes décrivant une date, nous constatons qu'il existe une forme décrivant seulement l'année. Cette année peut contenir un ou plusieurs preuves externes dont ils peuvent précéder l'ENA comme « عام/ l'année » ou la suivre comme « هـ/ hégire » dans l'exemple (6). L'indicateur gauche permet d'ajouter un degré de certitude au nombre identifié. Entre autres, pour se rassurer qu'un nombre détecté est bien une année et ne pas un nombre quelconque. La présence de cet indicateur dépend de la nature de l'année.

(6) عام 1434 هـ

En 1934 hégire

(7) في 2004

En 2004

En (7), nous avons une novelle émergence d'une date respectant la forme déjà mentionnée dont l'élément central est l'année. Dans ce cas, la forme de cette date est identifiée via la préposition « *في/ en* ».

La forme suivante est une date qui contient le mois comme étant un élément central. Cette date est incomplète puisqu'elle est composée de deux éléments seulement. Nous trouvons le nom ou/et le numéro du jour et le mois ou bien le moi et l'année. Cette forme se représente généralement comme une partie intégrante dans une autre ENA. Cette forme d'ENA est symbolique donc elle peut être assignée à la catégorie événement (8), (10) ou un nom de lieu

(9). Elle peut décrire aussi une date indiquant dans son contexte un évènement comme en (11) qui est la journée internationale de la femme.

(8) ثورة 14جانفي

La révolution de 14 janvier

(9) ملعب 14 جانفي برادس

Stade 14 janvier de Rades

(10) يوم العيد 1 شوال

Aïd al-Fitr, le 1er chawwâl

(11) يوم 13 أوت

Le 13 aout

Durant l'analyse de notre corpus, nous rencontrons une autre forme décrivant une date complète. Cette forme est composée de tous les éléments nécessaires pour exprimer une date complète (le nom ou/et numéro du jour, le mois et l'année). Dans quelques articles inclus dans notre corpus d'étude, nous constatons l'utilisation des anciens numéros arabes pour décrire l'année et le numéro du jour (14). Concernant la détection, nous trouvons une ENA sans ou avec des indicateurs. Ces derniers précédant la forme courante sont reliés au premier élément (le nom ou/et numéro du jour) comme « يوم/ le » qui est à son tour suivi par un signe de ponctuation « : » dans (15).

(14) ٣١ آب ٢٠١١

31 août 2011

(15) يوم : 10 كانون الثاني 2010

Le 10 janvier 2010

(16) ربيع عام 1990 م

Le printemps de 1990

(17) صيف 2000

Eté 2000

La forme finale d'une date que nous avons identifiée est celle basée sur une saison. Cette forme est très utilisée comme une date symbolique dans notre corpus d'étude. Rappelons qu'une saison est une division d'année, marquée par le changement de climat, utilisée comme un indicateur de temps. Elle est toujours suivie par une année comme en (16) et (17).

Les dates peuvent apparaitre aussi dans la forme ordinaire, sans avoir un contexte bien déterminé, mais elles dépendent de l'écriture de différents pays arabes. Par exemple, dans les articles provenant des pays orientaux nous constatons que les mois syriaque et musulmans sont les plus utilisés. Par contre, les mois grégoriens sont utilisés d'une façon fréquente dans les pays

magrébins. D'ailleurs, au sein de cette union, il existe une différence aux niveaux des appellations des mois. En Tunisie, le mois d'août en arabe est « أوت/ aout » de même qu'en Algérie tandis qu'en Maroc, son appellation est « غشت/ aout ».

2.2. Catégorie Nom de personne

La catégorie Nom de personne est dédiée à représenter les différentes formes décrivant un nom de personne arabe. Cette variété de formes est liée aux pays d'origine, la religion, la culture, le niveau de formalité et la préférence personnelle. En général, un nom de personne arabe contient cinq parties ne suivant aucun ordre particulier : al-ism, al-kuniyah, al-nasab, allaqab et al-nisba [Shaalan, 2014]. La combinaison de ces cinq parties permet de construire un nom de personne quand elles sont regroupées au sein de la même ENA. Rappelons la signification de chaque partie en donnant des exemples à partir de notre corpus d'étude.

2.2.1. Al-ism

Al-ism est ce qu'on appelle le prénom. Ce prénom est le premier nom donné à une personne lors de sa naissance. Le prénom peut être masculin comme « عبد الله/ Abdullah ; عادل/ Adel ; حسين/Hussein » ou féminin comme « فاطمة/ Fatma ».

2.2.2. Al-kuniyah

Al-kuniyah est un élément utilisé comme une forme informelle pour s'adresser à quelqu'un par respect comme l'utilisation de « oncle » ou « tante ». Elle indique aussi que quelqu'un est le père ou la mère d'une personne particulière. Par exemple, « أم كلثوم/ Om Kalthoum » signifie la mère de kalthoum.

2.2.3. Al-nasab

Al-nasab est un élément décrivant un nom patronymique qui commence par un lien comme « بن/bin » ou « بنت/bint » signifiant respectivement « le fils de » ou « la fille de ». Cet élément suit directement Al-ism (prénom) comme « فهد بن عبد العزيز / Fahad ibn Abdul Aziz » qui désigne « Fahd le fils de Abdul-Aziz ». Il faut mentionner que l'existence de lien n'est pas toujours obligatoire.

2.2.4. Al-laqab

Al-laqab est définie comme une épithète qui est généralement religieuse ou descriptive. Prenons les deux exemples suivants : le mot « الرشيد / Al-Rashid » signifie « le bien guidé » et le mot « الفضل / Al-fadl » signifie « le proéminent ».

2.2.5. Al-nisba

Le dernier élément à citer est Al-nisba qui est semblable à ce que les occidentaux appellent un nom de famille très connu comme la dynastie « آل نهيان /Al Nahyane ».

2.2.6. Les formes de noms de personnes identifiées

En fait, nous identifions 24 formes décrivant les formes alternatives d'un nom de personnes. Les ENA suivantes sont des exemples illustrant quelques formes identifiées durant notre étude linguistique. Nous remarquons que l'ENA se compose au moins d'un parmi les cinq éléments déjà expliqués.

(18) الأستاذ فؤاد بك العادلي

Le Professeur Foued Bek Al-Adeli

(19) بهرام باشا بن مصطفى باشا بن عبد المعين

Berham Pasha Ben Mustafa Pasha Ben Abd Al-Moen

(20) عبد الفتاح أبو غدة

Abd Al-fattah Abu Ghoda

(21) بازل أبن الليث أبن بازل

Bazel Ben Al-layth Ben Bazel

Dans l'exemple (18), l'ENA ayant la catégorie nom de personne est précédé par un indicateur externe exprimant une profession. Cette ENA contient un autre indicateur interne « بك /bek » situé entre le prénom et le nom de la famille. il s'agit d'une civilité. Dans l'exemple (19), l'ENA est identifiée en tant que « al-nisba » puisqu'il existe le connecteur « بن /ben ». Cette ENA contient deux indicateurs internes similaires « باشا /Pasha » situés au milieu qui sont également des civilités. L'ENA dans (21) décrit également « al-nisba » quoique cette fois nous n'avons pas l'indicateur dedans. Le connecteur indiquant « al-nisba » est une variante typographique de celui présenté dans (19). L'exemple dans (20), décrivant la catégorie nom de personne prend la forme ordinaire : un prénom suivi par un laqab ayant la forme de kuniyah.

2.3. Catégorie Nom de lieu

Lors de l'exploration de notre corpus, nous considérons la catégorie Nom de lieu en tant qu'un nom propre ou une expression désignant un lieu. Il existe trois sous-catégories, appartenant à la catégorie Nom de lieu, qui apparaissent fréquemment qui sont les suivantes :Nom de lieu absolu, relatif et géographique.

2.3.1. Nom de lieu absolu

Un nom de lieu absolu est un emplacement défini dans son sens par un seul endroit. Les ENA ayant cette sous-catégorie apparaissent dans notre corpus d'étude sous forme d'un nom de pays,

de ville, de délégation et de région. Ces formes sont identifiées via des mots déclencheurs et des prépositions (22, 23) par ce qu'ils jouent un rôle d'un indicateur de lieu.

(22) في تونس

En Tunisie

(23) في أوروبا

En Europe

(24) محافظة البصرة

La ville de Basra

(25) مديرية مناخة

La municipalité de Manakhah

(26) قرية الشبطلية

La région de Al-Shabtiliyah

Les indicateurs précédant les ENA identifiées diffèrent d'un pays à autre. Dans notre corpus d'étude, nous distinguons un ensemble d'indicateurs représentant des noms de lieux absolus comme {محافظة, ولاية, قضاء, لواء , مدينة}, {معتمدية, عمادة, دير, مديرية}, {قرية, منطقة, إمارة } pour décrire respectivement les noms de ville, de délégation et de région.

2.3.2. Nom de lieu géographique

Un nom de lieu géographique se réfère à un point physique spécifique sur la terre. Toutes les formes associées à cette sous-catégorie apparaissent avec des caractéristiques géographiques spécifiques soit montagne soit hydronyme. Les formes d'ENA décrivant un hydronyme sont divisées en deux ensembles dont le premier regroupe les noms de rivière ou du lac et le second est consacré aux noms de mer.

(27) جبل الشعانبي

Montagne de Chambi

(28) البحر الأبيض المتوسط

La mer méditerranée

(29) بحيرة سد الروم

Le lac de barrage de Rom

Dans l'exemple (27), l'indicateur interne « جبل / montagne » décrivant une montagne est parmi un ensemble des indicateurs synonymes que nous avons collecté comme {جبل, مونتي, جبال , الجبل قمة, الجبال, تل, }. Il existe aussi des synonymes étrangers appartenant à cet ensemble comme le mot « مونتي /montagne ». Les exemples (28) et (29) décrivent des ENA exprimant respectivement un nom de mer et de lac. Concernant les noms de rivières, ces derniers peuvent être identifiés à travers « نهر /rivière » ou « وادي /rivière » et sa variante typographie « وادى ».

Les noms du lacs peuvent également être identifiés par d'autres indicateurs comme « سبخة /lac » et « عين / lac ».

2.3.3. Nom de lieu relatif

Un nom de lieu relatif est une place qui est spécifié en termes d'autres nom de lieu absolu comme les noms de bâtiments. En analysant le corpus, nous identifions 16 sous-catégories associées à nom de lieu relatif. Ces sous-catégories sont détectées à travers des mots déclencheurs faisant partie de l'ENA et pouvant être défini ou indéfini comme par exemple « المتحف /le musée » et « متحف /musée ».

Tableau 6. Sous-catégories associées à un nom de lieu relatif

Sous-catégorie	Exemple
Musée	المتحف المسيحي المبكر بقرطاج
	Musée Paléo-chrétien de Carthage
Palais	قلعة السلع
	Le palais de Sela
Mosquée	الجامع العمري
	La mosquée d'Omari
Eglise	الكنيسة المارونية
	L'église maronite
Bain publique	حمامات عفرا المعدنية
	Les chutes minérales d'Ofra
Salle	قاعة المؤتمرات
	Salle de conférences
Hôtel	فندق الرويال
	L'hôtel royal
Marché	سيتي مول
	Centre commercial
Stade	ملعب إستاد السَّلام الرياضيّ
	Le stade sportif de Al-salam
Aéroport	مطار عمان المدني
	Aéroport civil d'Amman
Théâtre	المسرح الدولي بالجزائر
	Le Théâtre international algérien
Hôpital	مستشفى الملكة علياء
	L'hôpital de la reine Alia
Jardin	حديقة العامرأت الطبيعية
	Le jardin naturel Amrat
Monument	سارية العلم الأردني
	La hampe de drapeau Jordanien
Tombe	كعب بن عمير الغفاري الصحابي ضريح
	Le tomb de sahabi Kaab bin Amir Al-ghafari
Rue	شارع الثقافة
	Rue de la culture

Le tableau 6 regroupe des exemples que nous choisissons pour illustrer une parmi les formes de chaque sous-catégorie décrivant un nom de lieu relatif. D'après ce tableau, nous constatons qu'une ENA assignée à une de ces sous-catégories peut avoir les constituants suivants : un adjectif (الدولي / international), syntagme nominal (العلم الأردني / drapeau jordanien), nom commun (الثقافة / culture) ou encore un nom de lieu absolu (قرطاج / Carthage). De plus, nous identifions diverses natures des indicateurs pour chaque sous catégories. Prenons l'exemple des noms d'hôpitaux, l'indicateur interne « المستشفى /hôpital » a un ancien synonyme qui est « البيمارستان /hôpital » et un synonyme régional qui est « المشفى /hôpital ». De même pour les noms d'hôtels tel que l'indicateur « فندق /hôtel » fait partie de la liste suivante {ديدمان, خان, سراي }. Pour les noms de tombe, nous avons identifiées 12 indicateurs مزار, الحضرة, الروضة, المشهد, مدافن, مقبرة, حوزة, زاوية, مقام, مشهد, ضريح, مرقد,.

2.4. Catégorie Organisation

Dans notre corpus d'étude, nous rencontrons de nombreuses formes d'ENA décrivant la catégorie Organisation. Tous les indicateurs que nous avons détectés font partie de l'ENA identifiée. De plus, nous l'avons utilisé également pour déduire la nature de l'organisation. Les noms d'organisation identifiés possèdent les natures suivantes : Entreprise, Administration, Média, Organisation culturelle ou politique ou éducative. Il est vrai qu'il existe des organisations qui peuvent apparaître sous la forme d'acronymes. Néanmoins, l'utilisation des acronymes est relativement rare dans notre corpus d'étude.

Figure 11. Sous-catégories associées à un nom d'organisation

La figure 11 décrit les différentes sous-catégories dans lesquelles une ENA peut être assignée. Les formes identifiées nous permettent de traiter l'agglutination lorsque deux noms communs sont liés par une conjonction, tels que « والتعليم التربية /éducation et enseignement » et d'analyser un syntagme adjectival, comme par exemple l'ENA « التونسية الوطنية التلفزة /télévision nationale

tunisienne », contenant une succession d'adjectifs définis dont le mot « التونسية » est un adjectif arabe exprimant « al-nisba ».

2.5. Catégorie Evènement

Un événement est une composition nominale qui peut avoir différentes formes. Cette catégorie peut également être composée de sous-catégories. En se basant sur notre corpus d'étude, nous constatons qu'un événement peut être imbriqué dans un nom de lieu, une date ou un nom de personne. En d'autres termes, une date peut faire référence à un événement qui se produit dans un lieu dont cet événement peut être lié à une personne spécifique. En fait, nous avons identifié 3 sous-catégories d'un évènement qui sont : événement politique, évènement culturel et évènement religieux.

2.5.1. Evènement politique

Un évènement politique est un fait avec des conséquences importantes. Cette sous-catégorie peut être liée à des révolutions, des guerres ou des conflits. Il peut également décrire les célébrations liées aux républiques comme des fêtes d'indépendance. Dans notre corpus, nous constatons qu'un événement politique peut se représenter à travers des imbrications d'autres ENA. Prenons l'exemple de l'ENA « سليم أبو سجن مجزرة /le massacre de la prison d'Abu Salim », celle-ci décrit un événement politique contenant un nom de lieu relatif « سجن أبو سليم/ la prison d'Abu Salim » dont il inclut à son tour un nom de personne « أبو سليم /Abu Salim ». Une ENA ayant cette sous-catégorie peut contenir aussi un nom de lieu absolu c'est le cas de l'ENA « شغب بريطانيا /les perturbations de la Grande-Bretagne » d'où nom de lieu absolu est « بريطانيا / la Grande-Bretagne ». Nous trouvons également qu'un évènement politique peut avoir une date dedans tel que l'ENA « ثورة 14 جانفي/la révolution du 14 janvier ». Il est possible d'avoir une imbrication d'évènements politiques comme l'ENA « ذكرى استشهاد محمد الدرة / La mémoire du martyre de Mohammed al-Dura » dont le deuxième événement est « استشهاد محمد الدرة / Le martyre de Mohammed al-Dura ».

De plus, nous rencontrons les ENA liées à un verbe d'action tel que le nom « احتجاجات /les manifestations » dérivé du verbe « احتج /se manifester ». En outre, une ENA identifiée décrivant un évènement politique peut s'agir d'un nom qui évoque intrinsèquement l'événement tel que « تحرير تونس /Libération de la Tunisie ». Une ENA peut également être représentée par une nomination métonymique. Cela signifie que cette ENA est un nom ou une expression nominale prenant en compte la nature de l'événement tel que la date « السبت 8/1/2011 م /Samedi 01/08/2011 ».

2.5.2. Evènement culturel

Un événement culturel peut désigner un festival ou des journées culturelles telles que des festivals de musique ou de film. Pour identifier cette sous-catégorie, nous analysons les mots déclencheurs qui précèdent les ENA associées. En fait, ces mots déclencheurs peuvent être définis ou non définis. Dans le premier cas, l'ENA décrivant un événement culturel peut commencer par ce syntagme : un mot déclencheur défini suivi par un adjectif défini comme « المهرجان الدولي/ le festival international ». Dans le second cas, les mots déclencheur peuvent être par exemple « ملتقى, جائزة, مهرجان /festival, prix, forum ». La sous-catégorie évènement culturel a une large connexion avec les noms de lieux relatifs. Pour cette raison, nous trouvons qu'un événement culturel peut être composé uniquement par un indicateur interne indéfini suivi par un nom de ville « مهرجان حلب /festival de Halab ». Dans notre corpus, un événement culturel peut être avoir dans sa composition un événement religieux tel que « مهرجان عيد الفطر / festival d'Eid al-Fitr » sachant que l'ENA « عيد الفطر / Eid al-Fitr » est une fête religieuse en l'islam.

2.5.3. Evènement religieux

Un événement religieux est un fait lié aux fêtes religieuses comme « عيد الفصح/la pâque » ou la naissance des prophètes comme « المولد النبوي الشريف/le Mawlid ». Nous identifions cette sous-catégorie qu'à travers les mots déclencheurs. Quelques mots déclencheurs sont suivis par des syntagmes nominaux tel que « ذكرى المولد النبوي الشريف/le Mawlid ». De plus, une ENA décrivant cette sous-catégorie peut être imbriquée à des noms de personnes. Dans ce cas, ces noms de personne sont spécifiques et peuvent être précédés par un mot déclencheur faisant partie de classe fonction religieuse (النبي /le prophète). Prenons l'exemple d'une ENA identifiée lors de notre exploration de corpus qui est « مولد النبي محمد بن عبد الله/ la naissance du prophète Mohamed le fils d'Abdu-Allah », celle-ci est composé d'un nom de personne « محمد بن عبد الله/ Mohammed le fils d'Abdu-Allah » dont il décrit « al-nasab ».

La catégorisation effectuée contient cinq catégories principales ayant au total 47 sous-catégories. Ces catégories peuvent se chevaucher pour exprimer une nouvelle ENA. Dans ce qui suit, nous discutons les imbrications possibles entre les catégories d'ENA qui apparaissent dans le corpus d'étude. Nous alimentons cette section par des exemples illustratifs.

3. Imbrication des ENA

Dans notre corpus, les ENA ayant respectivement les catégories évènement et nom de lieu ont une relation de composition avec celles ayant la catégorie date. Par exemple, un nom de lieu relatif, spécialement les noms de stade, peuvent avoir contenir une date relative telle que « ملعب 14 جانفي/Stade du 14 janvier ». Nous devons mentionner que la forme de cette date respecte les

formes déjà identifiées durant notre étude linguistique. En fait, l'imbrication d'ENA peut être entourée par un contexte gauche comme par exemple «ملعب 14 جانفي بنابل/Stade du 14 janvier de Nabeul » et « ملعب 14 جانفي بالكاف / Stade du 14 janvier de El-kef ». Les contextes gauches sont à leur tour des noms de lieux absolus et agglutinés.

Parfois, les ENA Date réfèrent à un évènement symbolique qui a eu lieu au passé. Prenons l'exemple de l'ENA « 14 جانفي/ Le 14 janvier » qui est associée la révolution tunisienne. Pareillement, les ENA ayant les catégories évènement et nom de lieu peuvent être des parties intégrantes des noms de personne comme «ملعب الطيب المهيري بصفاقس / Stade El-Taieb Mhiri de Sfax » d'où « الطيب المهيري /El-Taieb Mhiri » est un nom de personnalité célèbre originaire de la ville Sfax. Les noms de personne peuvent se composer aussi avec les noms d'organisation comme par exemple « مؤسسة العنود الخيرية/Fondation charitable de Al-Anoud » sachant que le nom de personne dans cette ENA est le prénom d'une princesse.

L'étude linguistique que nous faisons joue un rôle important non seulement dans l'identification des formes d'une ENA mais aussi dans l'établissement d'une typologie d'ENA développée et étendue. Cette typologie profite du processus de raffinement effectué sur toutes les catégories et qui a touché même les sous-catégories. Il faut mentionner qu'avoir une typologie claire et riche en termes de catégories est un préliminaire pour réaliser d'autres tâches comme l'extraction des RS reliant les.

4. Conclusion

Dans ce chapitre, nous avons présenté une étude linguistique pour l'identification des ENA consistant à analyser un corpus d'étude extrait à partir de la Wikipédia arabe. Ce corpus regroupe des articles que nous avons collectés à partir de divers pays arabes et ils décrivent différents styles de la langue arabe. En fait, notre corpus d'étude nous a permis d'exploiter la richesse de sa structure pour identifier plusieurs formes décrivant une ENA. Afin de cerner les limites de ces formes et de les délimiter, nous avons proposé une définition précise et claire pour détecter les ENA. Cette définition est déduite après une catégorisation approfondie. La catégorisation obtenue a subi également un raffinement de telle sorte le niveau de granularité devient très important. Après avoir fourni une catégorisation de valeur, nous avons proposé une hiérarchie regroupant cinq catégories principales ainsi que leurs sous-catégories associées. La hiérarchie proposée fait appel à trois sous-hiérarchies décrivant les sous-catégories liées respectivement aux noms de lieu absolus, relatifs et géographiques. En se basant sur cette hiérarchie, nous pouvons passer à l'étape suivante de notre étude linguistique qui est l'identification des RS reliant les ENA.

Chapitre 4 : Typologie des relations sémantiques

L'étude linguistique pour la recherche des RS reliant les EN est une tâche que permet d'analyser et de comprendre la structure des textes. En fait, les RS permettent non seulement de relier les EN entre eux mais aussi de relier également ces EN aux informations pertinentes ayant une valeur sémantique importante. En outre, la détermination des RS permet de structurer les informations sémantiques et de les ordonner afin que les textes dans lesquels elles se trouvent soient des sources d'alimentation de lexiques d'EN. De plus, lors de la recherche des RS certaines EN apparaissent dans leurs formes brutes vu qu'elles étaient ignorées par le processus de REN effectuée préalablement sur le texte. Dans ce cas, cette recherche des RS vise à maintenir ce processus de REN. L'identification des RS intervient alors pour remédier aux lacunes du processus déjà mentionné. L'analyse de segments pertinents dont ses extrémités possèdent des catégories d'ENA raffinées offre l'opportunité de deviner le type de chaque RS identifiée. De même, les types de RS forment une typologie importante surtout quand ils subissent un raffinement donnant naissance à des nouveaux sous-types. Dans cette optique, nous explorons un corpus Wikipédia arabe pour construire une typologie de types et sous-types de RS développée et exploitable par diverses applications de TAL.

Le chapitre présent se compose de quatre parties dont la première est dédiée à la présentation de la définition de RS retenue qui est la base de notre recherche des RS entre les ENA apparaissant dans notre corpus d'étude. D'ailleurs, la recherche de ces RS constitue la deuxième partie de ce chapitre dans lequel nous donnons la définition de chaque RS en présentant des exemples illustratifs. Dans la troisième partie, nous expliquons le résultat d'une recherche consistant à découvrir les RS reliant les ENA à des informations pertinentes non catégorisées. Dans cette partie, nous traitons la relation entre une ENA avec une autre ENA qui n'est pas reconnue et annotée durant le processus de REN. De plus, nous abordons également le cas d'une ENA reliée à un mot significatif. La quatrième partie est dédiée à la représentation de la typologie de types de RS établie à la suite d'une classification basée sur les catégories reliées.

1. Définition des RS retenue

Après l'étude des formes de RS apparaissant dans le corpus d'étude, nous constatons qu'une RS relie toujours deux ENA. De plus, la liaison sémantique relie ces ENA à travers des indicateurs précis comme une expression ou un nom commun. Cette liaison sémantique se fait aussi selon un contexte qui se répète dépendant de la structure des articles Wikipédia. Pour délimiter une RS, nous la considérons en tant que :

« Une RS est un lien sémantique binaire reliant une ENA à une autre ou à un mot significatif pertinent, cette RS peut être exprimée indirectement ou directement à travers une expression simple ou complexe. »

La définition présentée ci-dessus est générique et valable pour tous les types qu'une RS sémantique peut avoir. Pour cette raison, nous nous basons sur l'étude linguistique faite pour rechercher les RS afin de récapituler les définitions en 18 types. Dans ce qui suit, nous présentons le processus d'identification des RS entre les ENA.

2. Identification des RS reliant des ENA catégorisées

La recherche des RS est une phase de découverte des liaisons sémantiques pouvant connecter les ENA entre elles ou de relier une ENA à une information pertinente. Cette recherche s'effectue préalablement sur un corpus annoté en termes d'ENA. Dans ce contexte, nous explorons et analysons un corpus d'étude extrait de la Wikipédia arabe contenants des ENA annotées. Nous rappelons que nous avons utilisé le même corpus exploité dans le chapitre précédent. Commençons par la présentation de la première RS qui est la synonymie.

2.1. Synonymie

La synonymie est un type de RS décrivant une similarité sémantique entre deux ou plusieurs ENA dans une même langue. La synonymie peut relier deux ENA ayant la même catégorie évènement plus précisément la même sous-catégorie comme dans l'exemple (1).

(1) الثورة التونسية التي تعرف أيضا بثورة الحرية والكرامة

La révolution tunisienne, également connu comme la révolution de la liberté et de la dignité

Nous avons pu détecter cette forme à travers l'expression « تعرف أيضا/connu aussi ». Cette expression se compose d'un verbe « تعرف/connaitre » et d'un adverbe « أيضا/aussi » regroupés ensemble pour dire que la première ENA de la phrase signifie la deuxième. La préposition « ب/par » permettait aussi d'introduire l'ENA en tant qu'un synonyme de celle qui la précédait. D'ailleurs, plusieurs expressions se présentaient dans le même contexte telles que « المسمى /nommé », « او /ou », « المكنى/ surnommé », etc.

(2) القديس سمعان المعروف بسمعان العامودي

Saint Samaan connu comme Samaan al-Ammoudi

Dans l'exemple (2), la synonymie relie deux ENA de catégorie nom de personne. Ce type de RS s'exprime à travers l'expression « ب المعروف /est connu par » contenant la préposition

« ب/par » attachée à la deuxième ENA. La séparation de cette agglutination sera effectuée ultérieurement lors de processus de REN.

(3) المولد النبوي أو مولد رسول الله

L'anniversaire du Prophète ou l'anniversaire du prophète d'Allah

La synonymie dans l'exemple (3) relie deux ENA ayant la catégorie évènement religieux. Le mot « أو/ou » séparant ces deux ENA joue le rôle d'un indicateur de synonymie.

2.2. Méronymie

La méronymie est un type de RS décrivant une relation d'inclusion. Ce type revient à représenter une partie d'un tout. En effet, deux ENA sont reliées via une méronymie signifie qu'une ENA fait partie de l'autre.

(4) كمال الملاخ من القاهرة

Kamal Al-Mallakh du Caire

(5) وزارة الدفاع في مصر

Ministère de la défense d'Égypte

Les exemples (4) et (5) illustrent deux formes de RS de type méronymie reliant des ENA à travers une préposition. La préposition « من » relie entre un nom de personne ou un nom d'organisation politique avec une ENA décrivant un nom de lieu pouvant être absolu comme dans ce cas.

(6) مهرجان عوافي يقام سنويا في منطقة عوافي

Festival Awafi s'organise chaque année dans la zone Awafi

En fait, la nature de la préposition diffère en connectant les catégories afin d'exprimer la méronymie. Prenons l'exemple de la préposition « في/ dans », celle-ci permet de relier un nom organisation et un nom de lieu dans l'exemple (6). Généralement, ce type de proposition apparait avec des verbes significatifs comme « يقام / se dérouler » et « يقع / se situer ». D'autres indicateurs peuvent prédire la relation courante à savoir le verbe « ينتمي / appartenir ».

2.3. Accessibilité

L'accessibilité est un type de RS reliant que les noms de lieux pour exprimer une sorte d'inclusion. Ce type signifie qu'il est possible d'accéder à une place à partir d'une autre. L'accessibilité relie les ENA dont la deuxième catégorie est un nom de lieu absolu.

(7) جبال سوريا محافظة السويداء

Montagnes Syrie du gouvernorat Sweida

(8) جبل حزنة (...) بمنطقة الباحة

Montagne Hazna (…) dans la région Elbaha

Dans l'exemple (7), l'accessibilité relie indirectement un nom de montagne qui se trouve dans la ville « السويداء / Sweida ». Ce type de RS se détecte à travers le mot déclencheur associé au nom de ville qui est « محافظة/ gouvernorat ». Le mot déclencheur précédant le nom de lieu peut être attaché à une préposition comme le mot « بمنطقة / dans la région » dans l'exemple (8). Cette agglutination assure la liaison sémantique puisqu'elle indique bien l'accessibilité en introduisant le nom de lieu.

(9) صنعاء (اليمن)

Sanaa (Yémen)

Dans l'exemple (9), l'accessibilité peut s'exprimer dans un contexte précis qui possède la forme d'un nom de ville « صنعاء/ Sanaa » suivi par un nom de pays « اليمن / Yamen » entre parenthèses. Ce contexte se présente fréquemment dans les titres d'articles dédiés à la description des capitales de pays.

2.4. Fonctionnelle

Fonctionnelle est un type de RS décrivant un rôle fonctionnel entre deux ENA. Ce rôle peut prendre plusieurs formes identifiées. Ce type de RS peut admettre des sous-types comme « fondateur » ou « directeur » ou encore « fondateur-directeur » et « organisateur ». Le sous-type « fondateur » est valable pour un nom de personne ou encore une organisation tandis que le sous-type « directeur » ne concerne qu'une personne.

(10) مهرجان دمشق السينمائي الدولي(...) تأسيس(...) محمد شاهين

Festival international du film de Damas (...) Fondation (…) Mohammed Shaheen

Dans l'exemple (10), la RS fonctionnelle relie deux ENA de catégories évènement et nom de personne via le nom commun « تأسيس / fondation ». Cet indicateur fait partie d'une liste contenant par ses synonymes ayant diverses catégories grammaticales comme le verbe « أسّس / fonder » ou le nom commun « مؤسس / le fondateur ».

(11) تنظم وزارة الثقافة المغربية، بتعاون مع مؤسسة الحسن الثاني للمغاربة القاطنين

بالخارج صيف كل سنة مهرجان روافد

Le ministère marocain de la Culture, en collaboration avec la Fondation Hassan II pour les Marocains résidant à l'étranger, l'été de chaque année organisent le festival Rawafed

La RS fonctionnelle présentée dans l'exemple (11) inclue le sous-type « organisateur ». Dans cet exemple, les organisateurs sont « وزارة الثقافة المغربية / Le ministère marocain de la culture » et « مؤسسة الحسن الثاني للمغاربة القاطنين بالخارج/ la Fondation Hassan II pour les Marocains résidant à l'étranger » et l'indicateur de la RS courante est le verbe « تنظم/organiser ».

2.5. Proximité

Proximité est un type de RS définie dans son sens par un rapprochement des endroits. Ce type de RS relient les ENA ayant la même catégorie qui est un nom de lieu. Plusieurs expressions permettent de déclencher la proximité sous diverses formes.

(12) جازان جنوب غربي السعودية

Jazan au sud-ouest de l'Arabie Saoudite

Dans l'exemple (13), la de proximité relie la première ENA « جازان / Jazan » à la deuxième ENA « السعودية / Arabie Saoudite » qui sont des noms de lieu absolus. Ce type de RS est exprimé à travers l'indicateur sémantiques « جنوب غربي/ sud-ouest » signifiant une approximation à travers les directions.

(13) لواء إسكندرون (...) قرب مدينة كسب

Gouvernorat Iskandaroun (…) près de la ville Kassab

L'exemple (13) décrit une nouvelle forme d'expression jouant le rôle d'un indicateur exprimant la proximité. Cette expression commence par le mot d'approximation « قرب / prés » suivi par un mot déclencheur d'un nom de lieu « مدينة / ville ».

2.6. Appartenance

Appartenance est un type de RS signifiant qu'une ENA appartient au sens mathématique à une autre ENA. Une parmi les deux ENA reliée joue toujours le rôle de l'ensemble général.

(14) جبل نيحا أحد جبال سلسلة جبال لبنان الغربية

La montagne de Niha, l'une de l'ouest chaîne de montagnes du Liban

(15) جبل أعظُم أحد جبال المدينة المنورة

La montagne Odhom, l'une des montagnes de Medina

Les exemples (14) et (15) représentent deux segments dont ses extrémités contiennent deux ENA ayant la même sous-catégories « nom de montagne ». Ces exemples possèdent la même

structure de représentation qui exprime une sorte d'appartenance. Ce type de RS s'exprime à travers le mot déclencheur « جبال / les montagnes » qui signifie un ensemble d'arrivée. Par conséquent, l'indicateur « أحد / l'un de » signifie dire que la première ENA est une partie de la deuxième ENA.

(16) جبل قنديل أعلى قمة جبل حاج إبراهيم

Le mont de Kendil, le plus haut sommet de la montagne Haj Ibrahim

Dans l'exemple (16), la structure diffère de celle se trouvant dans les premiers exemples déjà illustrés pour le type courant. Le nouveau segment se compose d'un adjectif « أعلى / le plus haut » qui indique un superlatif permettant de comparer la première ENA à une série cité après le mot « قمة / sommet ». Autrement dit, le nom de montagne « إبراهيم حاج جبل / la montagne de Haj Ibrahim » inclut d'autres montagnes.

2.7. Date de naissance

Date de naissance est un type de RS décrivant une date symbolique à laquelle une personne est née. Ce type relie la catégorie nom de personne avec les différentes formes d'une date.

(17) فاتن شاهين (11 يوليو 1951)

Faten Chahin (11 juillet 1951)

L'exemple (17) décrit une forme de représentation de la RS qui exprime une Date de naissance. Dans cette forme, le type de RS est exprimé via les deux parenthèses qui entourent une date complète référent à la naissance d'une personne.

(18) فيلدا سمور (...) تاريخ الولادة 8 أكتوبر 1955

Filda Samour (…) Date de naissance 8 octobre 1955

La deuxième forme de représentation, illustrée dans l'exemple (18), montre que la date de naissance se situe après une séquence de mots qui la relie à la personne appropriée. Autrement dit, la forme courante apparaît grâce à l'indicateur « تاريخ الولادة / Date de naissance ». Cet indicateur peut être remplacé par le nom commun « الولادة / la naissance ».

2.8. Date de décès

Date de décès est un type de RS décrivant une date symbolique référent à un acte de décès. De même pour la RS précédente, ce type relie la catégorie nom de personne avec différentes formes d'apparition d'une date. Mais, la forme de la date est toujours incomplète. En outre, la date de décès se compose juste par un numéro de jour et le mois.

(19) الفنانة ذكرى (...) الوفاة 28 نوفمبر

L'actrice Dhekra (...) Décès 28 novembre

(20) بكر الشدي تاريخ الوفاة 5 أكتوبر

Baker Alshadi date de décès 5 Octobre

L'exemple (19) décrit une forme d'appariation de la RS date de décès. Au sein du segment analysé, cette RS s'exprime via un indicateur important qui est le nom commun « الوفاة / le décès ». Dans d'autres segment candidats, cet indicateur peut être précédé par le mot « تاريخ / la date » déclenchant une date symbolique comme dans l'exemple (20).

2.9. Année de fondation

Année de fondation est un type de RS indiquant une date symbolique de fondation d'une organisation ou d'une ville ou d'un bâtiment. Ce type de RS peut être détecté grâce à des indicateurs ayant différentes natures. De plus, la nature de l'indicateur dépend de la position d'apparition de cette RS. Autrement dit, si la RS se trouve au début de l'article analysé alors l'indicateur ne peut être qu'un nom commun. C'est le cas présenté dans l'exemple (18).

(21) مهرجان القطن (...) عام 1965م

Festival de coton (…) Année 1965

Dans l'exemple (21), l'indicateur « عام / année » est suivi d'une date contenant juste une année pour décrire la fondation d'un festival. Pour cette raison, le segment étudié contient dans la première extimité la catégorie évènement culturel et dans la deuxième une date relative.

(22) جامع العادلية (...) تاريخ البناء 963هـ

Mosquée Al-adiliya (…) Date de construction 963H

L'exemple (22) introduit une nouvelle forme de la RS année de fondation qui relie une nouvelle catégorie (nom de lieu relatif) et une date symbolique. L'indicateur dans cet exemple est un syntagme nominal « تاريخ البناء / date de construction » vu qu'il décrit la date de construction d'un bâtiment.

(23) بني جامع النبي هوري (...) في العام 1276

La mosquée prophète Hori construit (..) dans l'année 1276

Le segment dans l'exemple (23) est pris du milieu du texte analysé. Pour cette raison, la forme de la RS année de fondation est différente de celles présentées dans les exemples (21) et (22). Ce type de RS s'exprime grâce à l'existence du verbe « بني / construit », la préposition « في /

dans » suivi par le nom commun « العام / l'année ». Ces indicateurs permettent de relier un nom de mosquée à sa date de fondation.

2.10. Familiale

Familiale est un type de RS exprimant une liaison familiale entre les noms de personne. Autrement dit, ce type de RS indique qu'une personne appartient à la famille de l'autre personne. Familiale est une RS qui possède des sous-types appelé « parent », « fils-fille », « sœur-frère », « mariage » et « divorce ».

(24) دينا هارون (...) الزوج خلدون وليد

Dina Haroun (…) Le mari Khaldoun Walid

L'exemple (24) décrit le sous-type appelé « mariage » associé au type de RS familiale. Ce sous-type relie le nom d'une femme au nom de son mari à travers l'indicateur « الزوج / le marie ». Dans le cas inverse, cet indicateur se transforme en « الزوجة / la femme ».

(25) فهد باسم (...) والده المنتج باسم عبد الأمير

Fahad Basem (...) Son père le producteur Bassem Abdul Amir

Dans l'exemple (25), le segment illustré exprime le sous-type « parent » faisant partie de la RS familiale. L'indicateur dans ce cas est le nom commun agglutiné « والده/ son père ». Si la RS familiale indique la mère de la personne décrite alors l'indicateur est « والدته/ sa mère ». Généralement, ce sous-type se présente que pour les parents qui sont célèbres.

(26) ديمة الجندي طليقة من المخرج فراس دهني

Dima Al-jondi divorcée du réalisateur Firas Dohni

Le sous-type « divorce » se trouve dans l'exemple (26) via l'indicateur « طليقة/ divorcée » suivi par la préposition « من/ de ». Cette préposition permet donc d'introduire le nom de l'ancien mari de la personne mentionnée au début de segment analysé. Cet indicateur s'utilise aussi dans le cas inverse sous sa forme masculine « طليق/ divorcé ».

(27) إنجي شرف وهي ابنة اللواء أحمد سامح، ولها شقيقة من والدتها هي الممثلة الجزائرية (…) نغم فتوكي

Enji Sharaf est la fille du major général Ahmed Sameh et elle a une sœur de sa mère, qui est l'actrice algérienne Nagham Fetouki

Le segment de l'exemple (27) contient plusieurs sous-types associés à la RS familiale. Le premier sous-type est « Fille » exprimé à travers l'indicateur « ابنة/ fille de ». Par contre, le

deuxième sous-type est « sœur » exprimé par l'expression « ولها شقيقة/ qui a une sœur » suivi après par le pronom relatif « هي/ qui est » introduisant le nom de personne.

2.11. Origine

Origine est le type de RS décrivant l'origine d'une personne (d'où il est issu). Ce type de RS relie un nom de personne à un nom de lieu qui est toujours absolu. Cette RS s'établit à l'aide des indicateurs sémantiques comme les mots communs ou encore un verbe significatif.

(28) فاتن شعبان الدولة مصر

Faten Shaaban, originaire du pays d'Egypte

D'après l'exemple (28), la RS origine s'exprime via le nom commun « الدولة / Pays » qui signifie l'origine de la personne mentionnée. Cet indicateur se situe au niveau de l'info-box de l'article Wikipédia. De plus, il est toujours suivi par un nom de pays comme « مصر / Egypte ».

(29) ولد فاروق الرشيدي بمصر

Farouk Alrashidi né en Egypte

La RS origine illustrée dans le segment de l'exemple (29) est décrite à travers le verbe « ولد / est né ». Ce verbe introduit le nom de personne suivi par son pays d'origine. Les deux ENA sont en liaison à travers la préposition « ب / en ». Rappelons que cette préposition est protégée par une balise lors de processus de REN.

2.12. Equivalence de dates

Equivalence de dates est un type de RS décrivant une équivalence entre deux dates appartenant à des calendriers différents. Autrement dit, ce type de RS traite une sorte de synonymie entre les dates et une équivalence entre eux du point de vue temporel. Les ENA traitées sont toujours équipées par des indicateurs qui se situent au milieu du segment pertinent décrivant le type de RS courant.

(30) سنة 1342هـ الموافق 1923 م

Année hégirienne 1342, correspondant à l'année1923 géorgienne

L'exemple (30) montre l'existence d'une équivalence entre les deux dates dont elles décrivent les deux années « 1342هـ / 1342 hégirienne » et « 1923 م / 1923 géorgienne ». L'équivalence dans ce cas se présente grâce au nom commun « الموافق / correspond » qui joue le rôle d'un indicateur de ce type de RS.

(31) 751 ميلادية و1350 هجرية

L'année géorgienne 751 correspondant à l'année hégirienne 1350

Dans l'exemple (31), l'indicateur d'équivalence de date diffère de celui de l'exemple précédent. Le type de RS étudiée relie deux dates à travers la conjonction « و / et » pour introduire le synonyme de la première date dans un autre type de calendrier.

2.13. Date politique

Date politique est un type de RS qui décrit une date symbolique associé à un évènement politique. Cette RS relie deux ENA ayant respectivement la sous-catégorie événement politique et la catégorie nom de lieu avec toutes les sous-catégories associées.

(32) الاحتجاجات البحرينية 2011

Soulèvement bahreïni de 2011

Dans l'exemple (32), le type de RS courant relie l'évènement politique à la date de son déclenchement sans avoir un contexte précis. Cette RS se présente exactement dans les noms d'articles de la Wikipédia décrivant des évènements politiques. Cette date politique est formée seulement par l'année.

(33) الثورة التونسية (...) الزمان 18 ديسمبر 2010

La révolution tunisienne (…) Date 18 décembre 2010

(34) الأزمة السورية (...) التاريخ 15 مارس 2011

La crise syrienne (…) Date 15 mars 2011

Au contraire de l'exemple précédent, la RS date politique peut relier deux ENA à travers des indicateurs sémantiques comme le nom commun « الزمان / date » dans l'exemple (33) ou son synonyme « التاريخ / date » dans l'exemple (34).

2.14. Place politique

Place politique est un type de RS décrivant un nom de lieu symbolique associé à un évènement politique. Cette RS relie deux ENA ayant respectivement la sous-catégorie événement politique et la catégorie nom de lieu plus précisément la sous-catégorie nom de lieu absolu.

(35) الاحتجاجات البحرينية (...) المكان البحرين

Soulèvement bahreïni (…) Lieu Bahreïn

D'après les deux exemples (35), la RS place politique est exprimée à travers l'indicateur « المكان / lieu ». En fait, ce nom commun se présente toujours comme un déclencheur unique de ce type de RS et il se situe au niveau des informations pertinentes fournies dans un article extrait de la Wikipédia.

3. Recherche des RS reliant les ENA à des informations significatives

La recherche des RS reliant les ENA à des informations non catégorisées permet de maintenir et améliorer le processus de REN. Autrement dit, les segments importants décrivant les RS à extraire peuvent être incomplets au niveau de l'annotation des ENA. Dans ce cas, la recherche que nous voulons faire essaie de signaler les ENA rencontrées qui sont reconnues et annotées en erreurs ou encore non reconnues et non annotées.

3.1. RS reliant des ENA à des autres ENA non catégorisée

Les RS dans notre corpus d'étude peuvent relier une ENA qui appartient à la hiérarchie d'ENA proposée dans le chapitre précédent à une ENA non catégorisée. Dans ce cas, nous constatons que la recherche de ces RS aide en premier pas à dégager les catégories manquantes et les rajouter dans la hiérarchie d'ENA. Dans ce contexte, nous avons trouvé une RS appelée Age de personne.

Age de personne est un type de RS exprimant l'âge d'une personne. Ce type de RS relie une ENA ayant la catégorie nom de personne avec un nombre exprimant l'âge repéré à travers des mots déclencheurs précis.

(36) فردوس محمد (العمر: 55 سنة)

Ferdous Mohammed (Age : 55 ans)

L'exemple (36) représente la forme d'un segment très fréquent exprimant la RS âge de personne. Le contexte d'apparition de ce type de RS a la forme des parenthèses dont ils contiennent un nombre et un indicateur d'âge « سنة / ans ».

3.2. RS reliant des ENA à des mots pertinents

Les RS peuvent relier des ENA à des mots pertinents et significatifs. En fait, ce mot pertinent décrit le type de la RS détectée et il peut être rajouté en tant qu'un mot déclencheur d'une ENA ultérieurement. Dans cette optique nous avons détecté trois types qui sont *nationalité*, *profession* et *religion*.

3.2.1. Nationalité

« Nationalité » est un type de RS décrivant la relation ou l'appartenance d'une personne à un pays bien déterminé. Ce type relie toujours un nom de personne à un adjectif « al-nisba ».

(37) دينا موسوي وهي ممثلة ومذيعة عراقية أوكرانية

Dina Mawsemi est une actrice et animatrice ukraino-irakienne

L'exemple (37) décrit un segment exprimant ce type de RS déjà mentionnée. Ce segment contient une succession d'adjectifs présentant deux nationalités « أوكرانية عراقية /ukraino-

irakienne » référant à la première ENA. Ces adjectifs sont introduits par le pronom relatif « هي /qui est » attaché à une conjonction. Il faut mentionner aussi que les deux nationalités se présentent à la suite d'une conjonction d'adjectifs signifiant les professions de la même personne mentionnée au début du segment.

(38) الجزائرية نغم فتوكي

L'algérienne Nagham Fetouki

Dans l'exemple (38), l'indicateur de la RS nationalité peut être aussi le mot déclencheur identifiant le nom de personne. Autrement dit, l'adjectif « الجزائرية / algérienne » joue un rôle lors de la détection de l'ENA et l'extraction de l'information sémantique entre cet adjectif et l'ENA qui le suit.

(39) فضل العلفي مخرج يمني

Fadhel Al-Alfi réalisateur yéménite

Dans l'exemple (39), le type de RS courant se détecte à travers un syntagme adjectival dont l'adjectif « يمني / yéménite » exprime une nationalité. Le premier élément de ce syntagme est aussi un commun qui fournit une information sémantique pertinente pour la personne.

3.2.2. Religion

Religion est un type de RS décrivant l'appartenance religieuse d'une personne. En outre, ce type de RS relie un nom de personne à un adjectif relatif à une religion.

(40) دينا الياسري مغنية مسيحية

Dina Yaciri, chanteuse chrétienne

L'exemple (40) décrit la seule forme qui décrit cette RS pour toutes ses occurrences dans les textes analysés. Le segment étudié commence par l'ENA décrivant un nom de personne suivi par un nom commun (profession) et par un adjectif relatif à sa religion « مسيحية / chrétien ». La RS religion s'exprime directement sans l'intervention des indicateurs sémantiques.

3.2.3. Profession

Profession est un type de RS associant un lien sémantique entre un nom d'une personne et son métier. Ce type de RS permet d'indiquer la position sociale de cette personne. La RS profession peut être exprimé à travers plusieurs formes d'indicateurs selon son contexte.

(41) الممثلة سالفة عويشق

L'actrice Soulafa Aouishek

Dans l'exemple (41), la RS profession est exprimée à travers le mot déclencheur « الممثلة / l'actrice » qui participe également à la reconnaissance de l'ENA étudiée. D'ailleurs, ce mot déclencheur fait partie d'une liste de nom commun exprimant une profession à savoir « المخرج, المهندس, المنتج ».

(42) بية الزردي هي مذيعة وممثلة تلفزيونية تونسية

Beya Alzerda qui est une animatrice et une actrice de télévision tunisienne.

Dans l'exemple (42), le segment exprimant la RS profession contient une conjonction de deux adjectifs « مذيعة / animatrice » et « ممثلة / actrice » référant à la même personne. Ces indicateurs sémantiques sont introduits par un pronom relatif « هي / qui est ». La RS de type profession apparaît fréquemment avec le type de RS « Nationalité ».

La recherche des RS reliant les ENA a donné la naissance de dix-huit types de RS possédant de différentes formes et divers contextes d'apparition. Les RS recherchées peuvent relier des catégories ou des sous-catégories ayant la même ou différente nature. De même, certaines RS déjà illustrées ont profité des autres informations sémantiques comme les noms communs significatifs, les adjectifs pertinents et encore des nombres apparaissant dans un contexte précis.

4. Hiérarchie de types de RS établie

La hiérarchie de types de RS que nous créons comporte trois classes principales contenant au total dix-huit types de RS. La classification des RS se fait à la base de la nature des ENA reliées.

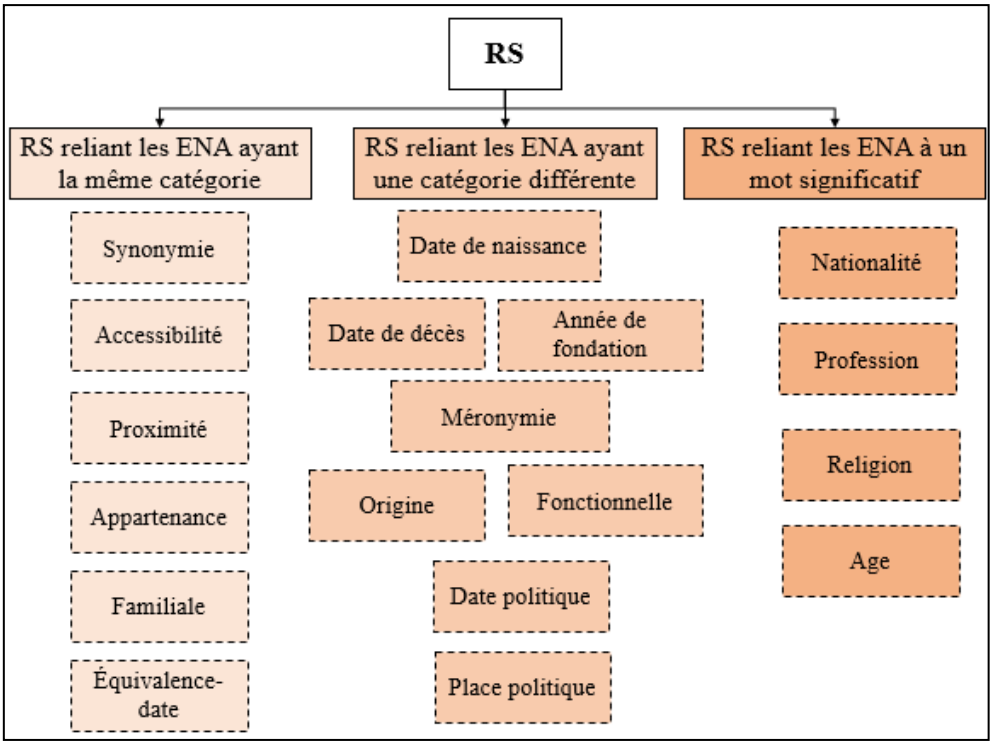


Figure 12. Hiérarchie de types de RS établie

La figure 12 illustre trois classes présentées dans cette hiérarchie dont la première comporte six types de RS reliant les ENA qui ont la même catégorie. Les segments pertinents, qui représentent ces six RS, sont entourés par une catégorie parmi les catégories : nom de lieu, nom de personne, date, organisation et événement. La deuxième classe contient 8 types de RS qui peuvent relier les ENA ayant des catégories différentes. Il faut mentionner que les catégories reliées peuvent se situer sur la même ligne ou dispersées sur plusieurs lignes. Finalement, la hiérarchie contient une troisième classe qui regroupe quatre types de RS reliant l'ENA à un mot pertinent qui lui ajoute une information sémantique importante.

5. Conclusion

L'étude linguistique que nous avons effectuée permet de chercher les formes d'apparition des RS à travers l'analyse des segments pertinents. Les segments sont tellement significatifs de sorte que nous avons pu identifier dix-huit types de RS en nous basant sur des définitions sémantiques précises pour chaque type. La richesse du corpus annoté et exploité nous a aidé à raffiner les types de RS pour qu'ils admettent des sous-types ce qui favorise la précision des informations sémantiques associées aux ENA reliés. L'étude linguistique que nous avons faite peut participer également à maintenir le processus de REN appliqué sur le corpus. En outre, les RS entre les ENA à une autre non catégorisée permettent de rajouter cette catégorie à la hiérarchie citée dans la première section. La hiérarchie de type de RS joue un rôle important pour la détermination du type d'information sémantique reliant les ENA. Les types représentés permettent de faciliter l'annotation de la RS identifiée ultérieurement. De plus, les formes alternatives apparaissant dans le corpus d'étude ont permis aussi de préciser une définition de RS qui participent à tracer les chemins d'extraction performants.

Partie 3 : Démarche proposée

Chapitre 5 : Démarche proposée pour la reconnaissance des entités nommées arabes

Après avoir effectué une étude linguistique profonde sur un corpus extrait du volume arabe de la Wikipédia, nous proposons une démarche de reconnaissance et d'annotation des ENA. Cette démarche profite de la définition d'ENA et de la catégorisation que nous avons établies précédemment. En outre, nous nous référons à la forte granularité offerte par la catégorisation pour détailler les éléments composant les ENA et les annoter. Pour ce faire, nous exploitons le formalisme des transducteurs pour réunir la reconnaissance et l'extraction, ce qui agit sur la réduction de temps d'exécution ultérieurement. De plus, les transducteurs sont caractérisés par leur rapidité et robustesse. Les transducteurs que nous allons concevoir exploitent les techniques de la plateforme Unitex. De ce fait, nous manipulons les ENA via des transducteurs d'analyse, de filtrage et de généralisation d'étiquetage. D'ailleurs, nous effectuons une analyse syntagmatique également. L'établissement de l'ensemble des transducteurs nécessite la modélisation des dictionnaires qui vont guider le processus de reconnaissance. Quant à l'annotation des ENA, ce processus va se baser sur une étape de normalisation de leur annotation des ENA afin de la conformer à la norme TEI.

Dans le chapitre courant, nous commençons par la description des étapes composant la démarche proposée pour la reconnaissance et l'annotation des ENA. Par la suite, nous présentons la partie concernant la modélisation de dictionnaires à utiliser pour effectuer la REN. Ensuite, nous décrivons les grammaires proposées pour les mots déclencheurs et les expressions régulières en illustrant des exemples. Puis, nous nous présentons les transducteurs établis pour traiter les particularités de la langue arabe comme l'analyse syntagmatique et la résolution de l'agglutination. Après, nous élaborons les transducteurs de reconnaissance et d'annotation des ENA pour les phases d'analyse, de filtrage et de généralisation d'étiquetage.

1. Démarche proposée pour reconnaître et annoter les ENA

La démarche que nous proposons pour reconnaître et annoter les ENA s'articule autour de deux étapes principales. La première étape est consacrée à la création des règles d'extraction et la création ou l'enrichissement de dictionnaires participant à améliorer la REN. La deuxième étape de cette démarche est dédiée à établir divers ensembles de transducteurs dont chacun possède un rôle bien déterminé. Les ensembles de transducteurs se partagent comme suit : des transducteurs traitant les particularités de la langue arabe, des transducteurs pour la reconnaissance et l'annotation des ENA et des transducteurs pour la normalisation. Le deuxième ensemble de transducteurs assurant la reconnaissance et l'annotation des ENA se décompose à son tour en trois sous-ensembles : les transducteurs d'analyse, de filtrage et de

généralisation d'étiquetage. Pour décrire les étapes de la démarche proposée, nous proposons le schéma illustré ci-dessous.

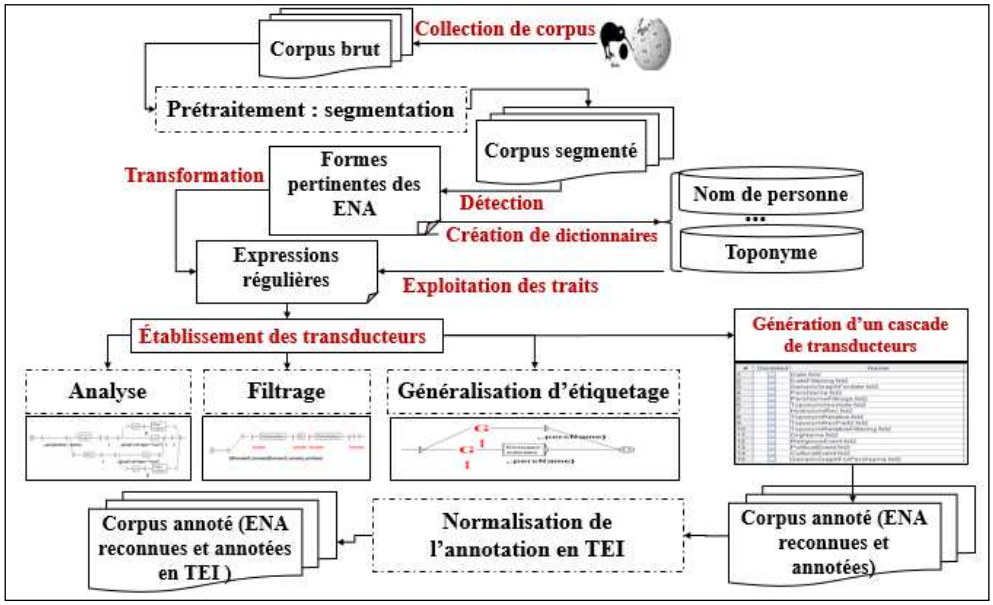


Figure 13. Etapes de la démarche proposée pour reconnaître et annoter les ENA

La figure 13 montre que la reconnaissance et l'annotation des ENA passent par plusieurs phases. Tout d'abord, il faut traiter les spécificités de la langue arabe qui consiste à résoudre les phénomènes linguistiques pouvant empêcher la reconnaissance. De plus, ce traitement permet d'effectuer une analyse syntagmatique ce qui favorise la création des grammaires locales pour qu'elles soient appelées au sein des transducteurs de reconnaissance et d'annotation des ENA. Ces grammaires locales permettent aussi de minimiser la taille de ces transducteurs déjà mentionnés pour ne pas avoir des chemins redondants. La deuxième classification des transducteurs donne naissance à des transducteurs d'analyse, de filtrage et de généralisation d'étiquetage. Cette classification ne se fait pas d'une façon aléatoire, mais elle est bien organisée vu que nous visons à réaliser une bonne REN. En fait, les traducteurs d'analyse sont contrôlés par ceux de filtrage pour corriger les erreurs obtenues. Les transducteurs de généralisation d'étiquetage exploitent la sortie de transducteurs d'analyse pour augmenter le champ de la REN et détecter les ENA apparaissant hors de leur contexte. Pour les transducteurs de normalisation, cet ensemble est dédié à transformer les annotations déjà faites à des balises associées à la TEI. Comme le montre la figure 13, le processus de reconnaissance et d'annotation d'ENA repose, entre autres, sur les dictionnaires pour identifier les chemins à parcourir. Pour cette raison, nous essayons de créer des dictionnaires ou d'enrichir ceux qui existent à partir des entrées collectées dans le corpus d'étude. Dans ce qui suit, nous commençons par la modélisation des dictionnaires nécessaires à exploiter ultérieurement.

2. Modélisation des dictionnaires

La création des ressources nécessaires joue le rôle d'un pont reliant l'étude linguistique faite sur le corpus d'étude et la démarche proposée. En d'autres termes, cette étape est le noyau de la démarche que nous proposons. Elle se compose de deux phases importantes participant à assurer une bonne reconnaissance des ENA. Ces phases sont la construction ou l'enrichissement des dictionnaires et l'identification des règles d'extraction. Les dictionnaires que nous essayons de modéliser ne possèdent pas la même structure. Réellement, cette modélisation dépend en premier lieu de la nature grammaticale de l'entrée à stocker et du niveau de la granularité sémantique que nous voulons atteindre en second lieu. Plus précisément, les informations à stocker dans les dictionnaires sont les suivantes : la catégorie grammaticale de l'entrée (nom, adjectif) et le trait sémantique qui définit le type (prénom, nom de la famille, etc.). Nous pouvons ajouter des informations additionnelles telles que le genre (féminin ou masculin) et le nombre (singulier, duel ou pluriel).

Pour les informations sémantiques à insérer dans les dictionnaires, nous effectuons un processus de mise en correspondance manuel afin d'obtenir un trait sémantique compréhensible. Dans ce contexte, nous proposons la liste suivante qui décrit les noms de dictionnaires correspondant aux catégories exploitées et les traits associés à leurs entrées.

Tableau 7. Mise en correspondance des traits de dictionnaires pour la REN

Catégorie	**Nature d'entrée**	**Trait associé**
Nom de lieu	Nom de continent	Np+Cont
	Nom de pays	Np+Country
	Nom de ville	Np+City
	Nom de délégation	Np+Delegation
	Nom de région	Np+Region
Nom de personne	Prénom	Np+Hum+FistName
	Nom de famille	Np+Hum+LastName
Date	Mois syriaque	Np+SyriacMonth
	Mois grégorien	Np+GregorianMonth
	Mois musulman	Np+MuslmanMonth
	Mois mauritanien	Np+MauritanianMonth
	Nom de saison	Np+Season
	Nom de jour	Np+Day

Le tableau 7 illustre un processus de mappage effectué manuellement pour attribuer à chaque entrée des dictionnaires un trait grammatical et sémantique. Le trait grammatical ne se crée pas arbitrairement. En fait, celui-ci doit respecter une liste de codes fournis par la plateforme linguistique que nous allons exploiter. Pour cette raison, le premier élément de trait est le code « Np » signifiant un nom propre. Le trait sémantique est un choix basé sur une certaine logique. Dans notre cas, nous choisissons un trait sémantique plus proche du nom de catégorie. D'après le tableau illustré, nous remarquons que trois catégories principales nécessitent de spécifier des dictionnaires de noms propres. Pour celles qui restent, elles peuvent se baser sur les mêmes traits mentionnés et sur d'autres dictionnaires comme celui des noms communs, d'adjectifs et de prépositions. Dans ce qui suit, nous allons présenter des grammaires pour les mots déclencheurs et les catégories d'ENA.

3. Grammaires proposées

La grammaire est une représentation formelle permettant d'organiser des éléments selon un ordre logique. Pour cette raison, nous essayons dans cette section de créer des grammaires pour organiser les mots déclencheurs ce qui permet de faciliter l'annotation des ENA ultérieurement. Nous créons également une grammaire pour arranger les mots déclencheurs, les catégories et les sous-catégories associées aux ENA. La grammaire que nous proposons vise à ordonner les chemins de reconnaissance et à les optimiser pour éviter les chemins redondants.

3.1 Grammaire pour les mots déclencheurs

Les mots déclencheurs sont des expressions jouant un rôle crucial lors de la détection des ENA. D'ailleurs, ils peuvent juste annoncer la présence des ENA ou être une partie intégrante de ces ENA. En fait, la classification des mots déclencheurs collectés permet de deviner la catégorie ou la sous-catégorie d'appartenance des ENA. Dans ce qui suit, nous présentons deux grammaires établies dont la première est dédiée à la catégorie nom de personne. La deuxième grammaire décrit les mots déclencheurs associés à la catégorie nom d'organisation.

3.1.1 Mots déclencheurs pour les noms de personne

Rappelons que plusieurs formes dérivant les noms de personne apparaissent dans notre corpus d'étude. Cette catégorie peut être identifiée avec ou sans mots déclencheurs sachant que les mots déclencheurs diffèrent selon plusieurs classes d'appartenance. Pour cette raison, nous décrivons la classification de mots déclencheurs que nous faisons sous forme d'une grammaire organisant les règles de production.

Mot déclencheur → Fonction artistique | Civilités | Fonction militaire | Fonction nobiliaire | Fonction politique | Profession | Fonction religieuse | Fonction sportive

Fonction artistique → المؤلف / l'auteur| المبدع /le créateur | الممثل /l'acteur| …

Civilités → السيد /M. | الآنسة /Mlle | السيدة /Mme. | …

Fonction militaire → الرائد /le major | الزعيم /le chef| المقدم /the colonel |…

Fonction nobiliaire → الأمير /le prince| الأميرة /la princesse| السلطان /le sultan| …

Fonction politique → الرئيس /le président| الوزير /the ministre| السياسي /le politicien| …

Profession → المعلم /le maitre| المدير /le directeur| الأستاذ /le professeur| …

Fonction religieuse → الإمام / l'imam| المؤذن /le muezzin | الرسول /le prophète|…

Fonction sportive → اللاعب /le joueur | الحكم /l'arbitre| …

La classification des mots déclencheurs nous a donné 8 classes contenant au total 277 mots. Certains mots déclencheurs peuvent se présenter au milieu de l'ENA. Prenons l'exemple d'un mot déclencheur « باشا / Pacha » appartenant à la classe Civilités se trouvant dans l'ENA suivante « بهرام باشا بن مصطفى باشا بن عبد المعين/ Bahram Pacha Ben Mustapha Pacha Ben Abdul Moin ». Les classes identifiées permettent non seulement de cerner les ENA mais elles sont exploitables également à l'identification des RS entre elles.

3.1.2 Mots déclencheurs pour les noms d'organisation

Les noms d'organisations possèdent plusieurs formes précédées toujours par les mots déclencheurs. En fait, ces mots déclencheurs sont parties intégrantes des ENA ayant cette catégorie. De plus, ils diffèrent selon leur nature. Pour cette raison, nous décrivons la classification de mots déclencheurs que nous faisons sous forme d'une grammaire pour illustrer les règles de production.

Nom d'organisation → Entreprise| Administration| Média| Organisation culturelle| Organisation politique | Organisation éducative

Entreprise → شركة/ société | مؤسسة/ fondation |…

Administration → مصلحة/ service| إدارة/ administration | …

Média → قناة/ canal | جريدة/ journal | التلفزة/ télévision |…

Organisation culturelle → نادي/ équipe | دار/ maison | …

Organisation politique → نقابة/syndicat | حزب/ parti | …

Organisation éducative → مدرسة/ école | كلية/ faculté | معهد/ lycée |…

La classification des mots déclencheurs liés aux noms d'organisations, nous a donné 6 classes contenant au total 45 mots.

Les grammaires des mots déclencheurs déjà illustrés peuvent être exploitées comme des éléments dans les chemins de reconnaissance des ENA. De plus, elles peuvent aussi être utilisées dans d'autres applications comme l'extraction des RS. En outre, les mots déclencheurs possèdent parfois des significations importantes pour relier deux ENA.

3.2 Grammaire proposée pour les ENA

L'identification des règles d'extraction consiste à délimiter des motifs représentatifs pour former des patrons syntaxiques. Rappelons que les règles d'extraction identifiées sont initialement des expressions indicatives apparaissant dans notre corpus d'étude. Ces expressions subissent un processus d'abstraction permettant de former des règles composées d'un ensemble de constituants. Ces constituants peuvent avoir des traits grammaticaux et sémantiques. Pour faciliter la création des graphes, nous traduisons ces règles d'extraction sous la forme de règles de production.

D'après les formes d'ENA déjà identifiées, nous avons des règles de production dépendant de mots déclencheurs et d'autres qui n'en dépendent pas. D'après notre corpus d'étude, nous distinguons cinq classes de règles de production principales pour reconnaître les catégories suivantes : *Date, Nom de personne, Nom de lieu, Evénement* et *Nom d'organisation*. Tout d'abord, nous présentons des classes déjà mentionnées sous forme d'une grammaire formelle. Par la suite, nous expliquons quelques règles avec des exemples illustratifs.

Expression_REN → Date | Nom de personne| Nom de lieu| Evénement| Nom d'organisation

Date → Période | Siècle| Année | Date basée sur le mois | Date complète | Date basée sur la saison | Filtrage-Date | Générique-Date

Personne → Mot déclencheur Nom de personne| Nom de personne

Nom de personne → Complet| Al-Kuniya| Al-Nisba| Al-Nasab

Nom de lieu → Absolu| Relatif| Géographique

Evénement → Culturel| Politique| Religieux

Nom d'organisation → Entreprise| Administration| Média| Organisation culturelle| Organisation politique| Organisation éducative

L'identification des expressions régulières est une phase qui facilite l'établissement des transducteurs. En fait, une expression régulière prend la forme d'un patron syntaxique pour avoir un modèle formel. Chaque règle contient des constituants et elle peut être transformée sous forme d'un graphe qui sera transformé à son tour en un transducteur dépendant de la plateforme linguistique exploitée.

Tableau 8. Expressions régulières créées pour quelques catégories d'ENA

Catégorie/ Sous-catégorie	Expression régulière
Année	<MD>+<NB>+[م/ه/هـ]
	سنوات 1980
Période basée sur le siècle	<Prps>+<القرن >+<NB>+[م/ه/هـ]+<Prps>+ <القرن >+<NB>+[م/ه/ه]
	من القرن 11 م إلى القرن 17 م
Nom de personne complet	<FirstName>+<genName>+<LastName>
	سليمان الأول القانوني
Nom de personne (Al-Kuniyah)	<FirstName>+<Kuniya>
	عبد الفتاح أبو غدة
Nom de lieu géographique	<MD>+<Adj+color>
	الجبل الأحمر
Nom de lieu relatif	<MD>+<Nc>
	كنيسة العذراء
Evènement culturel	<MD>+<Adj>+<Prps>+<Nc>+<Adj>
	المهرجان الدولي للتراث الشفهي
Nom d'organisation	<MD>+<Adj>+<Adj+Nisba>
	التلفزة الوطنية التونسية

Le tableau 8 illustre quelques exemples d'expressions régulières que nous créons pour effectuer le processus de REN. Ces expressions correspondent aux expressions pertinentes extraites à partir de notre corpus d'étude. Les constituants de chaque expression régulière ont des significations précises propres à la plateforme que nous allons exploiter. Il existe des constituants propres à nos préférences comme <MD> qui désigne un mot déclencheur sachant que MD est un ensemble de mot déclencheurs. Prenons l'exemple d'un nom de lieu géographique (nom de montagne) l'ensemble MD est composé de {الجبل, جبل, قمة, etc.}.

4. Etablissement des transducteurs

L'établissement de transducteurs consiste à représenter graphiquement la spécification des expressions régulières déjà créées. Cette représentation n'est pas aléatoire car nous devons fixer un principe à suivre pour la création des boîtes décrivant un chemin de reconnaissance. Par conséquent, chaque transducteur contient les mots déclencheurs collectés, en particulier ceux ayant des chemins communs. Ajoutons que nous essayons toujours de séparer les chemins chevauchés car ils peuvent produire des ambiguïtés. À l'intérieur d'un transducteur créé, nous pouvons appeler d'autres transducteurs pour éviter la redondance. De plus, cet appel participe à la minimisation de la taille ce transducteur créé.

Les transducteurs permettent non seulement la description des chemins de détection d'une certaine séquence mais aussi ils rajoutent également à ces chemins l'annotation souhaitée. Dans ce cas, la notion de transducteur montre qu'il est un formalisme capable de coupler les deux

tâches de la détection et l'annotation. Réellement, cela va nous permettre de minimiser le temps d'exécution et d'avoir un résultat adéquat. Dans cette section, nous illustrons les types de transducteurs que nous avons conçus. Ces transducteurs ne jouent pas seulement le rôle des graphes mais ils possèdent des particularités qui permettent de bons résultats dans le travail que ne voulons réaliser. Nous commençons par présenter les transducteurs dédiés au traitement des particularités de la langue arabe.

4.1 Transducteurs traitant la langue arabe

L'établissement des transducteurs pour la tâche de REN ou l'extraction des RS exige le traitement des particularités de la langue arabe. En outre, les ENA peuvent contenir des syntagmes nominaux qui nécessitent une reconnaissance spécifique selon leur composition. De plus, une ENA peut se composer de composantes agglutinées. Pour cette raison, nous avons pensé à traiter les formes alternatives décrivant les particularités de cette langue et les traduire sous forme de transducteurs qui seront appelés à la suite, par d'autres transducteurs de la REN. Dans ce qui suit, nous donnons une explication détaillée de ce traitement.

4.1.1 Résolution de l'agglutination

Nous avons rencontré le phénomène d'agglutination dans diverses positions dans les chemins de REN ou encore pendant l'extraction. D'ailleurs, tous les syntagmes traités précédemment peuvent être agglutinés. Prenons l'exemple de « مهرجان فاس لطرب الملحون » qui décrit une ENA ayant la catégorie événement culturel. Dans cette ENA, le syntagme nominal « لطرب الملحون » est agglutiné car il est attaché à la préposition « ل ». Le phénomène d'agglutination se pose non seulement pour les prépositions (« للمعارضة ») qui s'attachent à un mot, mais il est également fortement lié à la présence d'une conjonction. D'ailleurs, en langue arabe la conjonction est toujours attachée au nom qui le suit. Dans notre corpus d'étude, nous avons rencontré des conjonctions attachées à un nom propre comme « وأحمد /et Ahmed » ou « ولبنان / et Liban », aux noms communs aussi comme « والتعليم/ et l'enseignement » dans l'ENA « وزارة التربية والتعليم/le ministère de l'éducation et l'enseignement » et aux adjectifs comme « والكرامة / et la dignité » dans l'ENA « ثورة الحرية والكرامة / la révolution de la liberté et la dignité ». En fait, pour l'agglutination des noms propres, nous les traitons à fur et à mesure de leur reconnaissance. Nous avons créé des transducteurs qui traitent l'agglutination des noms communs et des adjectifs. Dans la figure suivante nous illustrons le transducteur qui reconnaît une conjonction ou une préposition attachée à un nom commun.

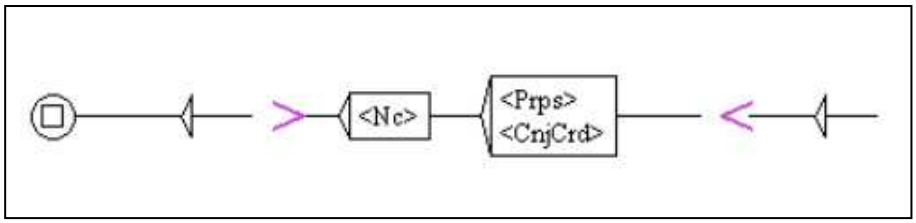


Figure 14. Transducteur résolvant l'agglutination dans le cas d'un nom commun

La figure 14 montre la forme d'un transducteur qui résout un cas d'agglutination apparaissant dans notre corpus d'étude. C'est le cas lorsqu'une préposition ou une conjonction est attachée à un nom commun. Les cases ayant la forme de < et > signifient que la reconnaissance est au niveau morphologique. Par conséquent, le transducteur reconnaît d'abord la préposition ou la conjonction et le nom commun en second lieu. Ce transducteur fonctionne si seulement si les deux dictionnaires des prépositions et des noms communs sont définis en tant que dictionnaires morphologiques dans la plateforme utilisée. D'ailleurs, il faut mentionner aussi que c'est grâce à cette plateforme que nous avons pu insérer les boîtes en violet décrivant le mode morphologique.

4.1.2 Analyse syntagmatique

Une ENA peut être composée par un syntagme nominal ou prépositionnel. Pour cette raison, nous avons créé des transducteurs traitant les différents arguments qu'un syntagme nominal ou prépositionnel peut avoir. Pour clarifier l'idée, nous avons commencé par la représentation de ces syntagmes et de leurs formes possibles apparaissant dans notre corpus d'étude.

Un syntagme nominal est une d'une succession de mots dont le premier possède toujours le trait grammatical « nom commun ». En outre, ce nom commun est la tête de ce syntagme pouvant être suivi par un autre nom commun pour exprimer un état construit. Par contre, un nom commun peut être suivi par un ou une succession d'adjectifs pour présenter un syntagme adjectival. Dans les deux figures suivantes, nous donnons les chemins de reconnaissance d'un syntagme nominal qui traite seulement les formes existantes dans notre corpus.

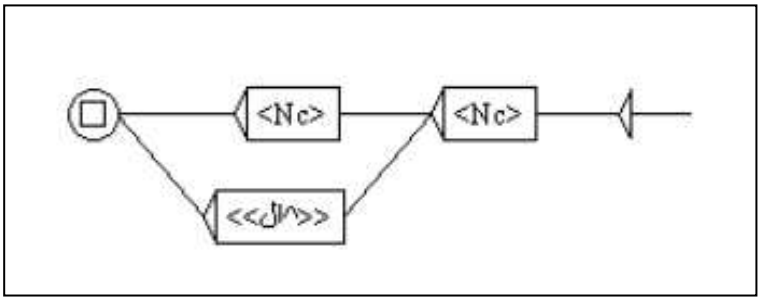


Figure 15. Transducteur traitant un état construit

La figure 15 décrit le transducteur reconnaissant un état construit. Ces deux chemins permettant de reconnaître deux noms communs successifs. Nous avons rajouté dans le second chemin la contrainte suivante : le deuxième nom doit être défini à travers la boîte contenant la valeur <ال^>. En fait, cette boîte désigne que nous sommes en train de faire une analyse morphologique à travers l'option « filtre morphologique » offerte par la plateforme exploitée. Les chemins illustrés sont appelés fréquemment ou bien dans un évènement culturel ou bien dans les noms d'organisation. Ce transducteur peut reconnaître le syntagme nominal « طلبة تونس/ les étudiants tunisiens » dans l'ENA « اتحاد طلبة تونس / Union des étudiants tunisiens » de catégorie Nom d'organisation politique. La contrainte d'avoir un nom défini peut être appliqué dans le syntagme suivant « نجباء المعهد/ les intelligents de l'institut » dans l'ENA qui est « نجباء المعهد مسابقة /La compétition des intelligents de l'institut » décrivant un nom d'évènement culturel. Rappelons que nous avons créé ce transducteur pour minimiser la taille des ceux qui en auront besoin dans leurs chemins de reconnaissance.

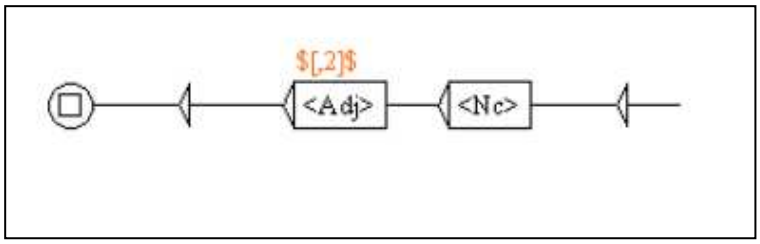


Figure 16. Transducteur traitant un syntagme adjectival

La figure 16 décrit le transducteur traitant un syntagme adjectival. Ce dernier se compose initialement par un nom commun décrit par le trait <Nc>. Ce nom commun peut précéder qu'un ou deux adjectifs au maximum. Cette contrainte est fixée suite à notre analyse à toutes les formes décrivant un syntagme adjectival dans le corpus d'étude. Graphiquement, la contrainte déjà citée est décrite à travers la deuxième boîte plus précisément l'intervalle au-dessus de celle-là. Cet intervalle est une sorte de condition qui se fixe en écrivant $[,2]$. Dans la sortie de la boîte courante. Plusieurs catégories ont besoin d'appeler ce transducteur comme les noms d'organisation. Par exemple, le syntagme « الثقافة المغربية /la culture marocaine » qui peut être précédé par un nom déclencheur exprimant un nom d'organisation politique pour construire l'ENA suivante « وزارة الثقافة المغربية /le ministère de la culture marocaine ».

Un syntagme prépositionnel commence toujours par une préposition qui peut avoir différentes valeurs. Dans les différentes valeurs assignées à ce syntagme, nous n'avons rencontré que le problème de l'agglutination. Autrement dit, les prépositions ne sont attachées à aucun composant qui les précède.

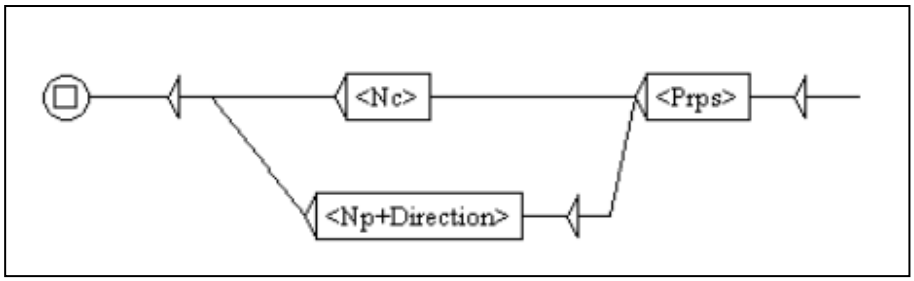


Figure 17. Transducteur traitant un syntagme prépositionnel

La figure 17 montre la forme d'un transducteur traitant un syntagme prépositionnel. Ce transducteur sera très utilisé fréquemment durant la REN. En outre, le processus de reconnaissance repose sur des patrons syntaxiques dans lesquels nous pouvons rencontrer un mot déclencheur précédé par une préposition. Par exemple, le segment d'extraction « في مدينة / dans la ville » sera détecté par le premier chemin. En fait, le syntagme prépositionnel peut contenir également un nom propre indiquant une direction dans le chemin de reconnaissance comme « في الجنوب /dans le sud ».

4.2 Transducteurs de reconnaissance et annotation des ENA

Les transducteurs que nous élaborons pour reconnaître et annoter les ENA respectent les règles d'extraction déjà construites. Nous distinguons trois ensembles de transducteurs qui sont tout d'abord des transducteurs d'analyses assurant la REN, puis ceux de filtrage qui permettent de filtrer les ENA qui ont été ignorées lors du passage des transducteurs d'analyse. Le troisième ensemble de transducteurs repose sur une nouvelle technique dédiée à la généralisation d'étiquetage. Tous les ensembles déjà mentionnés visent à fournir une bonne reconnaissance des ENA. Commençons par l'explication des transducteurs d'analyse.

4.2.1 Transducteurs d'analyse

Les traducteurs d'analyse permettent de tester plusieurs types de boîtes pour identifier les catégories et les sous-catégories qu'une ENA peut avoir. Au sein de transducteurs d'analyse, nous avons besoin de réutiliser les graphes traitant la langue arabe que nous avons présentés précédemment. Certains cas exigent la création d'un graphe local pour résoudre un petit problème d'agglutination. Dans ce qui suit, nous illustrons quelques transducteurs d'analyse pour monter le principe de leur établissement.

Pour la catégorie *Date*, nous créons un transducteur principal qui regroupe tous les autres sous-transducteurs. Chaque sous-transducteur est dédié à la reconnaissance les formes décrites et expliquées dans l'étude linguistique. Parmi ces transducteurs, nous choisissions de représenter celui qui reconnaît une saison suivie par une année.

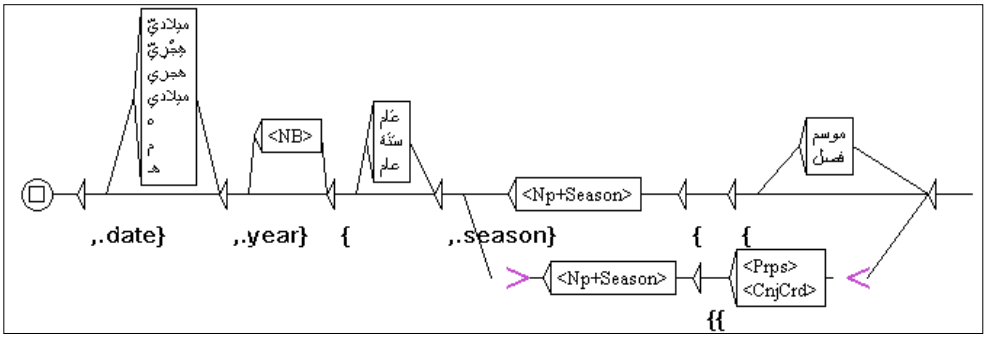


Figure 18. Reconnaissance une date composée par une saison et une année

La figure 18 montre les chemins alternatifs qui reconnaissent une date composée par une saison suivie par une année. Nous constatons que ce transducteur résout localement un problème d'agglutination lié à un trait de dictionnaire. Autrement, ce graphe permet de reconnaître une saison qui est stockée dans le dictionnaire et attachée à une préposition ou encore à une conjonction. Parmi les ENA qui peuvent être reconnues et annotées par ce transducteur, nous citons « 2000 صيف/ l'été 2000 » et « م 1990 ربيع عام/ le printemps de l'année 1990 ». Lorsque le chemin de reconnaissance débute par un mot déclencheur, ce transducteur est capable de reconnaître par exemple « فصل الخريف/ la saison d'automne ».

Il est très important de noter que nos transducteurs d'analyse agissent sur le texte par des annotations particulières. Ces annotations se basent sur les accolades en tant que marqueurs. En effet, le choix de genre de marqueurs n'est pas aléatoire car ils permettent de transformer l'ENA en un mot polylexical. Autrement dit, l'ENA reconnue et annotée ne peut jamais être détectée et touchée par un autre transducteur. Cela signifie que cette ENA est protégée. Il faut dire aussi que les marqueurs d'annotation se présentent que dans la sortie d'une boîte. Nous suivons ce principe d'annotation pour traiter le reste des catégories.

Pour la catégorie *nom de personne,* nous créons deux transducteurs principaux dont le premier transducteur appelle les sous-transducteurs traitant cette catégorie sans avoir des mots déclencheurs. Par contre, le deuxième sous-transducteur fait appel à une collection des mots déclencheurs avant les chemins de reconnaissance. Rappelons que les mots déclencheurs de la catégorie nom de personne ont été classés en 8 classes. En fait, l'utilisation des mots déclencheurs permet de réduire le taux d'ambiguïté que nous pouvons avoir.

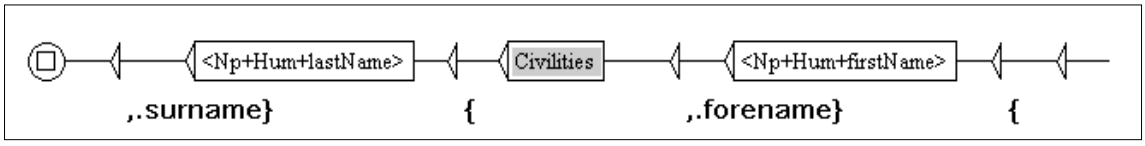


Figure 19. Transducteur reconnaissant un nom de personne

La figure 19 décrit un chemin de reconnaissance d'une forme d'ENA ayant la catégorie *nom de personne.* Cette ENA est composée par *un prénom* et un *nom de famille* séparé par une civilité comme par exemple « فؤاد بك العادلي/ Foued bek Al-Adly » qui contient une civilité faisant partie des civilités dans le sous-graphe en boîte grise. Ce transducteur ne contient pas une annotation finale « persName » car celui-ci sera appelé dans un transducteur d'analyse principal.

Dans les transducteurs précédents, nous n'avons pas utilisé des attributs à l'intérieur des balises vu que nous traitons juste des catégories principales. Par contre, nous allons utiliser un attribut appelé « type » pour stocker les valeurs possibles prises par une sous-catégorie. Afin d'illustrer l'utilité de cet attribut, nous commençons par la description d'un transducteur reconnaissant la catégorie nom de lieu et plus précisément un nom de lieu absolu.

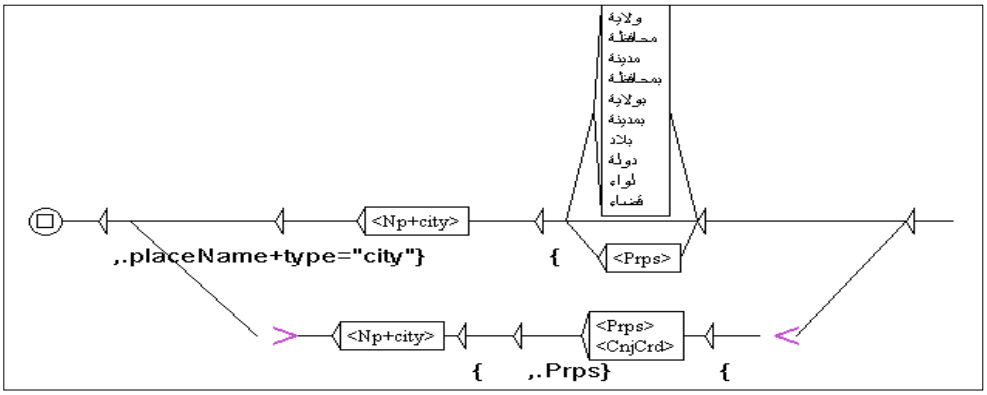


Figure 20. Transducteur reconnaissant un nom de ville

La figure 20 décrit la manière de reconnaissance d'un nom de ville qui va être annoté via la balise « placeName+type="city" ». Le signe « + » dans la balise n'est qu'une protection de celle-ci pour ne pas mettre un espace et générer des problèmes d'ambigüités ultérieurement. Dans ce transducteur, l'attribut « type » aide à déterminer la valeur de la sous-catégorie qui est un nom de lieu absolu. Cet attribut peut avoir les valeurs suivantes "country", "continent", "delegation" et "region" en dépendant de la nature de ce nom de lieu absolu. L'agglutination est traitée par ce transducteur pour protéger les prépositions et la conjonction. Cette protection est justifiée au fait qu'un nom de lieu absolu peut être une partie intégrante d'une autre ENA. Dans le cas échéant, cette dernière ne peut pas être reconnue.

C'est vrai que l'attribut « type » est utilisable principalement pour décrire une sous-catégorie, mais nous pouvons en profiter pour décrire une caractéristique géographique. Cet attribut, nommé encore élément, nous permet de spécifier la nature d'un nom de lieu géographique et les hydronymes.

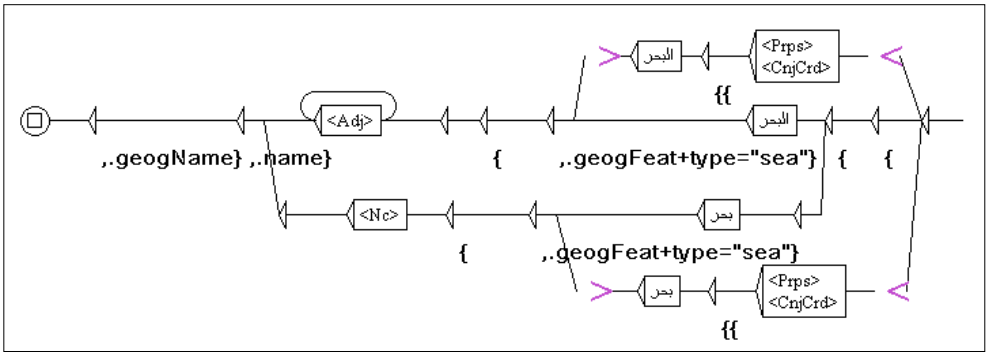


Figure 21. Transducteur reconnaissant un nom de mer

La figure 21 montre la manière d'annotation d'une sous-catégorie d'un hydronyme plus spécialement les noms de mer. Au début de chaque chemin, nous considérons les mots déclencheurs comme des caractéristiques géographiques associées à une balise appelée « geogFeat ». Par la suite, nous entourons le reste de l'ENA par une balise appelée « geogName ». Le transducteur illustré est désigné par un transducteur principal qui lui ajoute une balise globale appelée « placeName » contenant à son tour un élément « type="Hydronym" ».

Cependant, nous constatons qu'il existe des ENA non reconnues par la liste des transducteurs établis. Ces ENA sont non reconnues malgré l'existence de leurs chemins de reconnaissance. Pour cette raison, nous nous comptons effectuer une phase de filtrage pour reconnaître et annoter les ENA ignorées.

4.2.2 Transducteurs de filtrage

Les traducteurs de filtrage sont les fruits d'une phase de filtrage pour récupérer les ENA ignorées par les transducteurs d'analyse. Nous avons essayé de diagnostiquer ce problème et nous avons constaté qu'il est causé par la phase de prétraitement. Autrement dit, en segmentant les textes, les composants de certaines ENA sont séparés par une balise de segmentation.

Nous essayons de créer de nouveaux chemins de reconnaissance et d'annotation d'ENA en rajoutant la notion de variables. Ces derniers jouent un rôle très important parce qu'ils vont récupérer le contenu des boîtes de transducteur courant et les organiser au nœud final. Il faut mentionner donc que l'annotation dans un transducteur de filtrage diffère de celui d'un transducteur d'analyse. Dans la figure suivante, nous présentons un transducteur de filtrage permettant de reconnaître une ENA ayant la catégorie nom de personne.

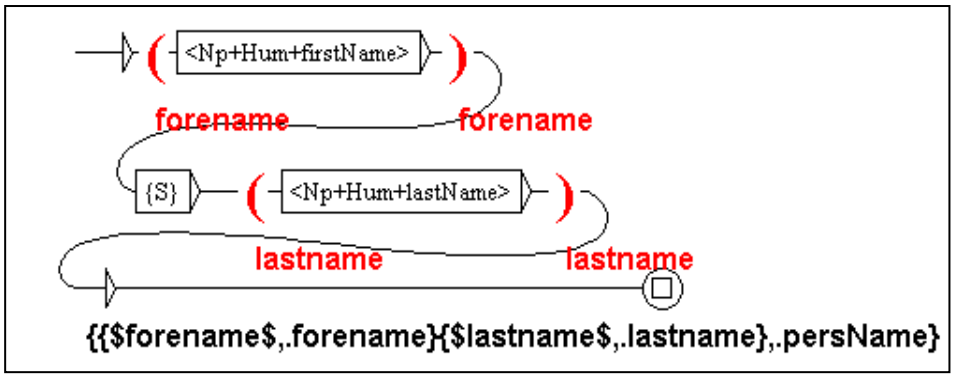


Figure 22. Transducteur de filtrage reconnaissant un nom de personne

La figure 22 décrit un transducteur de filtrage traitant un cas exceptionnel de la reconnaissance d'un nom de personne. Dans ce cas, le chemin de reconnaissance est séparé par une balise de segmentation. Ce symbole a interrompu la reconnaissance de l'ENA bien que le même chemin ait été défini dans un transducteur d'analyse. Les variables utilisées sont marquées en couleurs rouges dont elles sont appelées au dernier nœud. En fait, leur appel est assuré par le symbole $.

La phase de filtrage n'a pas touché seulement la catégorie *nom de personne* mais elle a remédié le même problème posé dans d'autres catégories. En réalité, l'établissement de ces deux ensembles de transducteurs reste encore insuffisant vu que nous avons encore des ENA non reconnues. Dans ce qui suit, nous décrivons une nouvelle stratégie pour détecter les formes d'ENA non traitées.

4.2.3 Transducteurs de généralisation d'étiquetage

Les transducteurs de généralisation d'étiquetage sont des graphes visant à localiser l'occurrence des ENA non reconnues lorsqu'elles apparaissent hors de leurs contextes définis. La création de ce genre de transducteur consiste à établir un chemin dont la première boîte est représentée par un grand G en gras et une accolade ouvrante. De plus, ce chemin possède une deuxième boîte contenant la catégorie recherchée afin de retrouver l'ENA non reconnue. L'application d'un transducteur de généralisation d'étiquetage ne s'effectue pas d'une façon ordinaire comme les transducteurs déjà illustrés.

En fait, le transducteur d'étiquetage générique récupère toutes les ENA existant dans le fichier texte déjà mentionné et ayant la catégorie recherchée. Puis, ce transducteur revient au texte pour chercher ces ENA et les annoter lorsqu'elles apparaissent hors de leurs contextes. Par exemple, le transducteur d'étiquetage générique peut reconnaître un prénom (« جميلة/ Jamila ») en se basant sur un nom de personne compet (« جميلة التونسي/ Jamila Al-Tounsi ») et l'annoter en tant que nom de personne. Il est évident que à travers ce genre de transducteurs nous n'allons pas poser les problèmes habituels pour un prénom d'une personne ; est ce que

« جميلة/ Jamila » est un adjectif ou un prénom. A l'intérieur d'un transducteur d'étiquetage générique, nous pouvons ajouter des restrictions sur la boîte contenant la catégorie recherchée plus précisément sur l'annotation de sortie. Notamment, nous pouvons enrichir l'annotation de sortie par des informations supplémentaires. Dans la figure suivante, nous illustrons les restrictions que nous avons faites pour reconnaître seulement un prénom ou un nom de famille apparaissant hors de leur contexte.

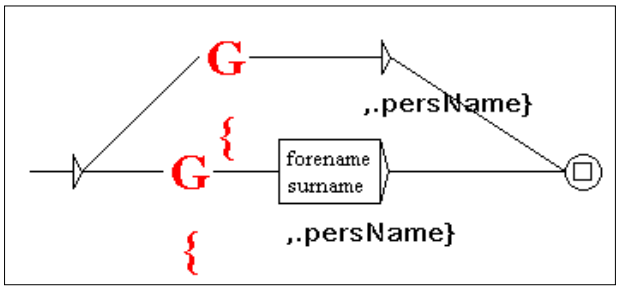


Figure 23. Transducteur de généralisation d'étiquetage pour les noms de personnes

La figure 23 illustre un transducteur de généralisation de graphe pour traiter les ENA ayant la catégorie *nom de personne.* En fait, nous dupliquons le nœud contenant un grand G pour chaque chemin sinon le chemin de reconnaissance ne sera pas fonctionnel. Dans ce cas, la restriction est faite dans le deuxième chemin.

Dans notre corpus d'étude, nous constatons aussi que les mois et les années apparaissent fréquemment hors de leur contexte. Ces éléments non reconnus, pouvant composer une date complète ou équiper par des mots déclencheurs précis, participe à diminuer la précision d'un système de REN.

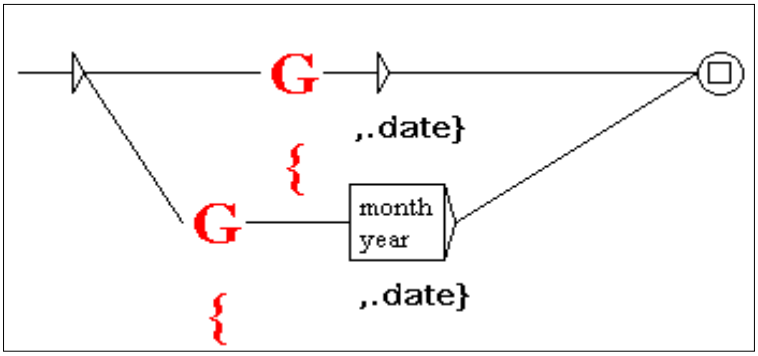


Figure 24. Graphe d'étiquetage générique pour la catégorie Date

La figure 24 décrit le transducteur d'étiquetage générique dédié à la catégorie *Date*. Dans ce graphe, nous essayons de récupérer les dates non détectées hors de leur contexte. De plus, nous utilisons la restriction pour les mois et les années comme par exemple l'année « 2018 » qui a été étiquetée via le mot déclencheur « سنة / année ».

L'établissement de transducteurs de REN se base sur plusieurs phases dont chacune essaie de remédier aux lacunes de ses précédentes. De plus, chaque phase possède ses particularités qui, une fois réunis, forment un système de REN performant. Les transducteurs d'analyse ont joué un rôle important pour fixer les chemins assurant une bonne reconnaissance. Ensuite, ces transducteurs regroupent des annotations structurées et détaillées. Les transducteurs de filtrage ont participé à améliorer le champ de reconnaissance en se basant sur le premier ensemble de transducteurs établis. En outre, ils ont réussi à récupérer des ENA non reconnues précédemment. La puissance des transducteurs d'étiquetage générique nous a permis de résoudre le problème d'occurrence des ENA hors contextes en se basant sur celles ayant des contextes précis. De même, ces transducteurs nous ont permis d'éviter les problèmes d'ambiguïtés exposés par plusieurs chercheurs. En général, les traducteurs que nous avons construits se basaient sur diverses options à savoir la notion du mode ou filtre morphologique, la notion de variable, l'aspect générique. Dans ce qui suit, nous continuons à les exploiter en rajoutant d'autres options fournies également par la plateforme exploitée.

4.3 Transducteurs de normalisation de l'annotation des ENA

Les transducteurs de normalisation ou encore de synthèse consistent à transformer les annotations faites durant la phase de reconnaissance des ENA en celles liées à la norme TEI. Rappelons que la TEI est un consortium international dont l'objectif est le développement d'un standard permettant de préparer et échanger les textes électroniques. En conséquence, nous établissons nos transducteurs de normalisation en suivant la syntaxe TEI.

Nous rappelons que la syntaxe TEI est définie comme suit : une balise ouvrante décrivant comme par exemple <persName> et une balise fermante </persName>, ces deux balises entourent l'ENA. La balise <persName> peut inclure une imbrication d'un prénom, nom de famille et un mot déclencheur qui peut précéder un nom de personne. Ces éléments se trouvent respectivement au sein des balises suivantes <forename>, <surname> et <roleName>. Dans la dernière balise, il est possible de spécifier le type d'un mot déclencheur comme par exemple une fonction sportive ou une profession. Toutes les catégories et les sous-catégories que nous avons identifiées précédemment peuvent être représentées par la norme TEI. Par exemple, nous utilisons <placeName> et </placeName> pour décrire un nom de lieu et cette balise peut avoir un élément « type » et une valeur associée comme type="castle" pour les noms de châteaux dont ils font partie des noms de lieux relatifs. Pour mieux comprendre le principe de normalisation, nous proposons l'architecture suivante.

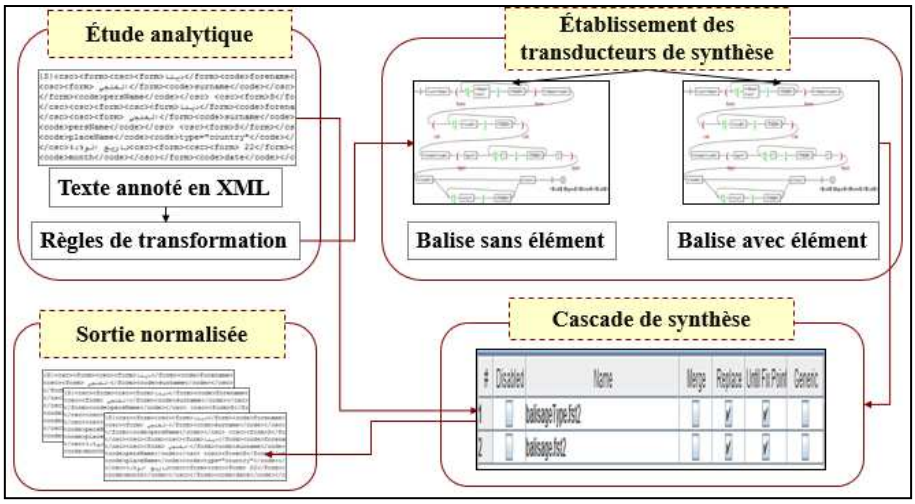


Figure 25. Architecture illustrative de la normalisation de l'annotation d'ENA

La figure 25 montre le principe que nous fixons pour normaliser l'annotation des ENA. En fait, nous commençons par une étude analytique consistant à consulter le fichier texte contenant les ENA reconnues et annotées. Ce fichier possède deux versions dont la première est annotée selon les accolades déjà expliquées. La deuxième est une version XML associée à la première version. Pour la normalisation, nous utilisons la deuxième version pour dégager des règles de transformation. Deux cas peuvent se présenter qui sont : le traitement des balises qui ne contiennent pas des éléments ou celles qui les englobent. De ce fait, nous obtenons deux transducteurs de normalisation. Dans la figure suivante, nous illustrons la forme du transducteur pour les ENA qui n'admettent pas un élément.

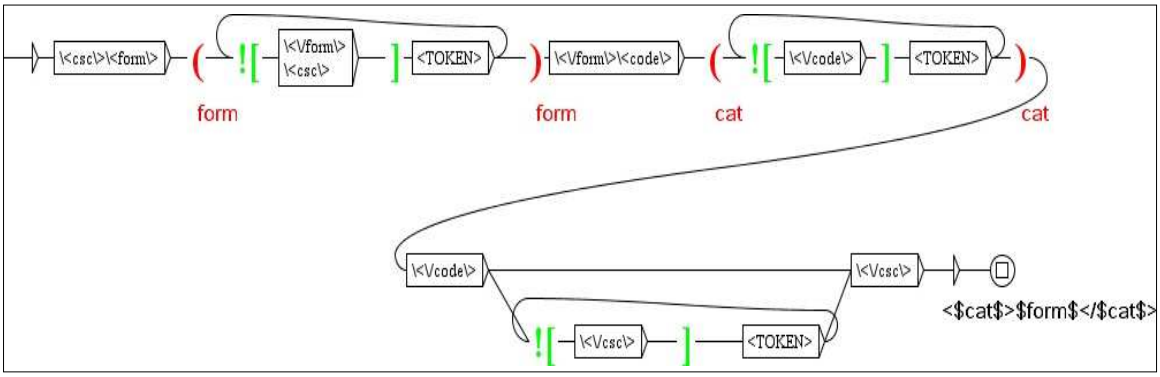


Figure 26. Normalisation de l'annotation d'une ENA sans élément

La figure 26 décrit le premier transducteur de synthèse sachant qu'il sera appliqué sur un texte balisé en XML, ce texte en XML est généré automatiquement par la plateforme que nous exploitons. Le transducteur illustré contient un chemin formé par plusieurs types de boîtes. Il existe des boîtes contenant l'annotation à transformer, c'est le cas de la première boîte par exemple. Il y a aussi des boîtes sous formes de variables (représentées en couleur rouge). Ces variables ont des noms significatifs parce qu'elles seraient exploitées dans l'annotation de

sortie. Il existe également des boîtes utilisant le concept de contexte négatif. La condition d'arrêt pour effectuer le traitement est décrite en dehors des boîtes **!** [,]. Rappelons que l'annotation de sortie se trouve dans la dernière boîte en respectant la syntaxe TEI.

Lors de la présence de l'élément « type », il y aura une petite modification dans le premier transducteur. Il s'agit de rajouter des boîtes après la variable appelée « cat » vu que le type toujours suit la catégorie principale.

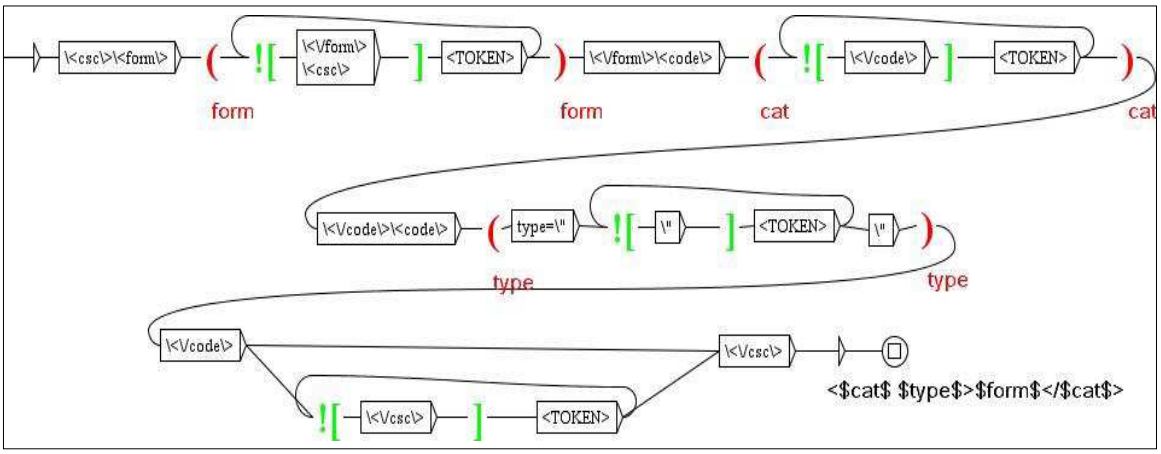


Figure 27. Normalisation de l'annotation d'une ENA sans élément

La figure 27 montre que nous avons adopté le même principe que dans le transducteur précédent mais nous avons traité la présence d'un élément dans une telle balise. Rappelons que le rôle de cet élément est de spécifier la sous-catégorie de l'ENA.

La visualisation des résultats de l'application de ces transducteurs sera plus claire dans la partie dédiée à l'implémentation. En fait, la tâche de transformation n'est pas une tâche facile car elle s'appuie sur la précision au niveau des chemins de reconnaissance. De plus, cette transformation se fait d'une façon itérative jusqu'à avoir toutes les ENA normalisées. En outre, la normalisation des ENA fonctionne selon un algorithme quoiqu'il soit représenté graphiquement à travers la notion des variables et des contextes. Cet algorithme admet une condition d'arrêt fixée dans la boîte entourée par le contexte négatif.

5. Conclusion

Dans ce chapitre, nous avons proposé une démarche de reconnaissance et d'annotation des ENA à partir d'un corpus Wikipédia arabe. Cette démarche se base principalement sur divers ensembles de transducteurs traduisant les expressions régulières établies. Les transducteurs conçus varient selon les tâches à réaliser l'analyse, l'annotation ou la normalisation. Nous avons élaboré des transducteurs pour résoudre le phénomène d'agglutination et d'autres transducteurs pour une analyse syntagmatique. Cette analyse s'occupe des ENA imbriqués contenant des syntagmes nominaux ou prépositionnels. Les transducteurs d'analyse se regroupaient en trois

ensembles d'analyse, de filtrage et de généralisation d'étiquetage ce qui favorise la qualité de reconnaissance. Les transducteurs de filtrage ont remédié au problème de segmentation qui découpe parfois les composants d'ENA. Les transducteurs généralisation d'étiquetage ont pu récupérer les ENA qui apparaissent hors de leur contexte. La conception de nos transducteurs s'appuie également sur des technologies avancées à savoir la notion de variables, le contexte négatif et le mode morphologique. La démarche proposée a inclus aussi une modélisation des dictionnaires que nous allons exploiter au sein de chaque transducteur.

Cependant, les transducteurs que nous avons conçus ont besoin d'un ordre de passage pour assurer le processus de REN. De plus, ils nécessitent un mode de passage pour appliquer les fonctionnalités de chaque technique exploitée. Pour cette raison, nous allons passer à leur implémentation en utilisant la plateforme linguistique Unitex.

Chapitre 6 : Démarche proposée pour l'extraction des relations sémantiques

Après avoir étudié notre corpus d'étude extrait à partir de la Wikipédia arabe, plus précisément de sa version issue d'un processus de REN, nous proposons une démarche d'extraction de RS entre les ENA basée sur l'approche symbolique. Cette démarche repose principalement sur la hiérarchie d'ENA que nous avons établie dans le troisième chapitre. Elle est considérée comme une partie complémentaire à la reconnaissance des ENA. Les segments pertinents reliant les ENA et exprimant une RS sont la base de l'établissement des expressions régulières. Ces segments participent également à créer un dictionnaire sémantique pour stocker les indicateurs sémantiques associées aux RS détectées. De plus, ce dictionnaire sémantique contient les entrées dégagées ainsi leurs traits significatifs. En fait, l'établissement des expressions régulières va faciliter la conception des transducteurs d'analyse qui comportent les chemins d'extraction et les annotations des RS. Au sein de chaque transducteur, nous avons besoin de structurer l'annotation des RS pour avoir une sortie structurée. Pour cette raison, nous allons faire recours à des fonctionnalités avancées de la plateforme Unitex telles que l'utilisation des variables et l'exploitation du mode morphologique pour traiter l'agglutination. De plus, l'extraction des RS entre ENA requiert le traitement de segments longs car les ENA peuvent se situer sur plusieurs lignes d'un texte arabe traité. L'annotation des RS se fait selon une balise respectant la norme TEI.

Dans ce chapitre, nous illustrons les étapes composant la démarche proposée à travers un schéma descriptif. Ensuite, nous passons à présenter chaque étape à travers des exemples explicatifs extraits du corpus d'étude préalablement annoté via un processus de REN. En outre, nous présentons la création des ressources nécessaires qui sont la construction d'un dictionnaire sémantique prioritaire et la conception des expressions régulières nécessaires. Après, nous passons à décrire l'établissement des transducteurs d'analyse, pour l'extraction et l'annotation des RS entre les ENA, relatifs aux expressions régulières. Dans cette partie, nous abordons le traitement des segments ayant des éléments agglutinés et la manipulation des chemins d'extraction longs.

1. Description de la démarche proposée pour extraire des RS

La démarche que nous préconisons pour l'extraction des RS reliant entre elles les ENA se compose de trois étapes : la modélisation d'un dictionnaire sémantique, la création des règles d'extraction et l'établissement des transducteurs. En fait, nous avons pu profiter de l'exploration et de l'analyse de notre corpus d'étude pour modéliser manuellement un dictionnaire sémantique permettant de guider le processus d'extraction des RS. L'identification des règles d'extraction se fait à travers l'étude des segments pertinents pour former des expressions régulières à transformer sous forme de graphes. L'ensemble des transducteurs établis seront

appelés au sein d'une cascade selon un ordre de passage prédéfini donnant un nombre minimal d'erreurs. Dans l'architecture suivante, nous décrivons les étapes de l'extraction des RS entre les ENA.

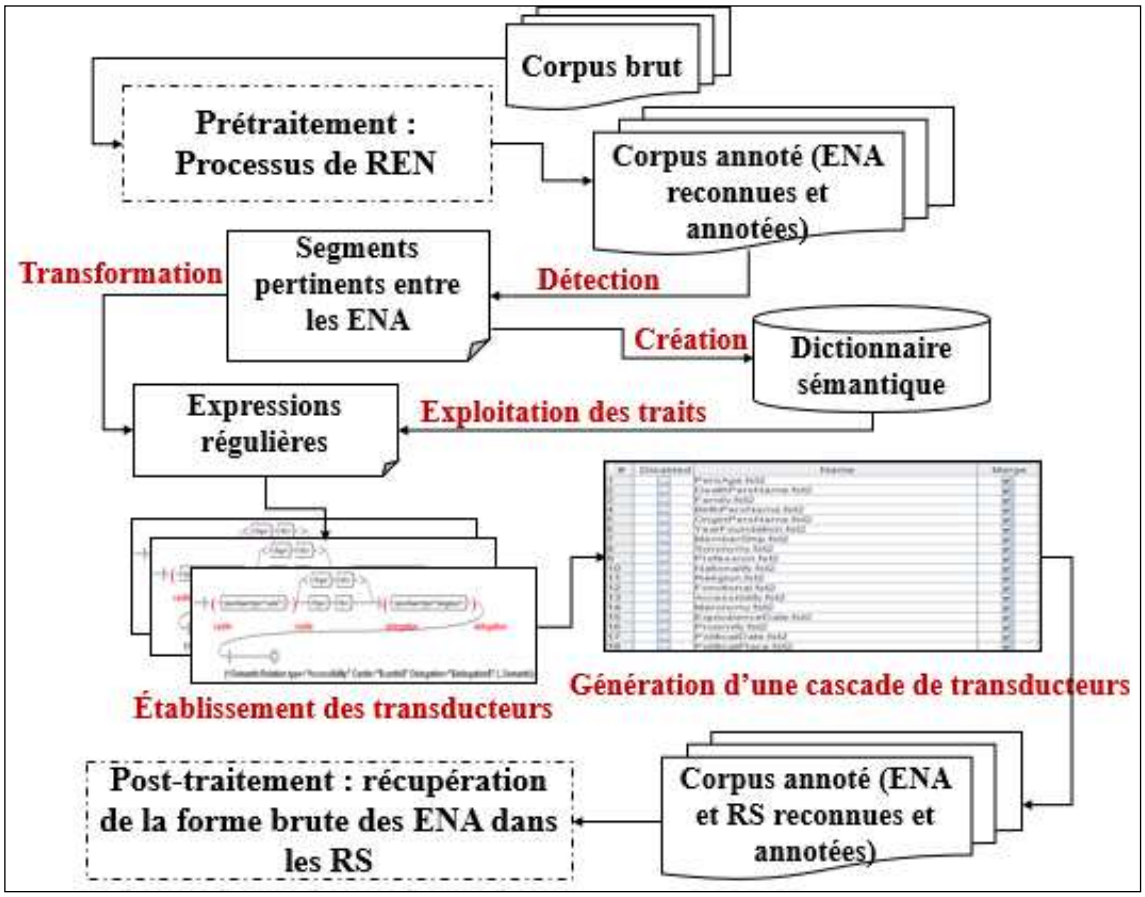


Figure 28. Processus d'extraction des RS entre les ENA

La figure 28 montre qu'à partir de notre corpus d'étude extrait de la Wikipédia arabe, nous avons identifié des règles d'extraction des RS en se basant sur les catégories associées ENA déjà reconnues et annotées par un système de REN. Le raffinement de la catégorisation nous a beaucoup aidé à envisager les difficultés rencontrées lors de l'extraction des ES. La richesse de la hiérarchie nous a permis de dégager des RS non traitées d'avance. Nous exploitons également des dictionnaires déjà créés et des dictionnaires sémantiques propres à ce processus d'extraction de RS. Les traits grammaticaux et sémantiques sont adoptés lors de la construction des règles d'extraction et leur transformation en des expressions régulières ainsi l'établissement des transducteurs d'extraction des RS. Les transducteurs établis doivent être regroupés au sein d'une cascade d'analyse dont son application se fait sur un corpus de test issu également d'un processus de REN.

Dans ce qui suit, nous expliquons chaque étape qui compose notre démarche avec plus de détails.

2. Etablissement des expressions régulières

La détection d'une RS reliant les ENA avec des règles d'extraction identifiées consiste à extraire toutes ses formes dans le corpus d'étude. Parmi les constituants d'une RS, nous trouvons les catégories déjà annotées dans le corpus exploité. Ces catégories sont utilisées comme des entrées lexicales. En fait, c'est l'avantage de l'annotation à travers les accolades qui rendent les ENA reconnues des mots polylexicaux jouant le rôle des traits grammaticaux.

Dans un traitement textuel, nous sommes toujours à la recherche de segments comportant un aspect spécifique. Ces segments possèdent une structure qui peut agir ultérieurement sur des textes pour extraire des RS identifiées. Il faut noter qu'une expression régulière est considérée comme un modèle défini permettant de caractériser un ensemble de chaînes de caractères. Pour cette raison, nous expliquons dans ce qui suit la transformation des règles d'extraction sous forme d'expressions régulières en donnant des exemples illustratifs extrait à partir de notre corpus d'étude.

2.1. Expression régulière associée à la RS Equivalence date

Nous rappelons que le type de RS « équivalence date » relie deux ENA ayant la même catégorie (date). En fait, ce type de RS permet de faire la correspondance entre deux dates appartenant à différents calendriers. Grâce à la diversité des articles composant notre corpus, nous avons identifié deux expressions régulières valables pour tous les types d'articles. Parmi ces expressions, nous choisissons d'analyser le segment décrivant la règle d'extraction identifiée qui sera présenté sous forme d'expression régulière.

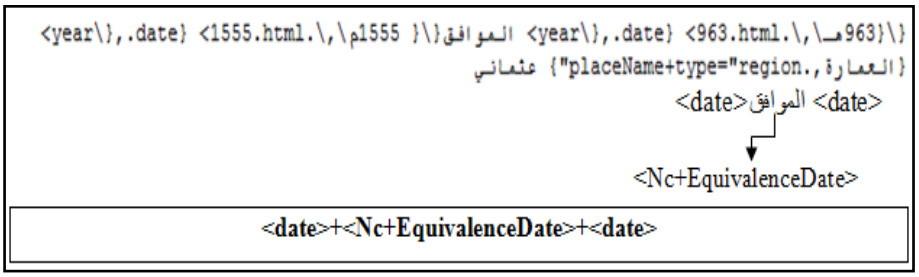


Figure 29. Expression régulière de la RS équivalence date

Dans la figure 29, le type de RS équivalence date est déduit à partir du mot déclencheur « الموافق/ équivalent à » qui joue le rôle d'un identificateur signifiant l'équivalence. Ce mot déclencheur sera stocké dans un dictionnaire sémantique avec le trait suivant : <Nc+EquivalenceDate> où « Nc » signifie un nom commun. A titre d'indication, nous rappelons que <date> est la catégorie des ENA reliées.

2.2. Expression régulière associée à la RS Date politique

Le type de RS date politique exprime une date symbolique associée à un déclenchement d'un évènement politique. Nous devons mentionner que celle-ci relie deux ENA ayant respectivement la sous-catégorie événement politique et la catégorie date. En fait, ce type de RS apparaît dans les articles contenant les évènements politiques. À la suite d'une analyse profonde, nous avons dégagé deux expressions régulières. A travers le segment suivant, nous présentons une expression régulière identifiée.

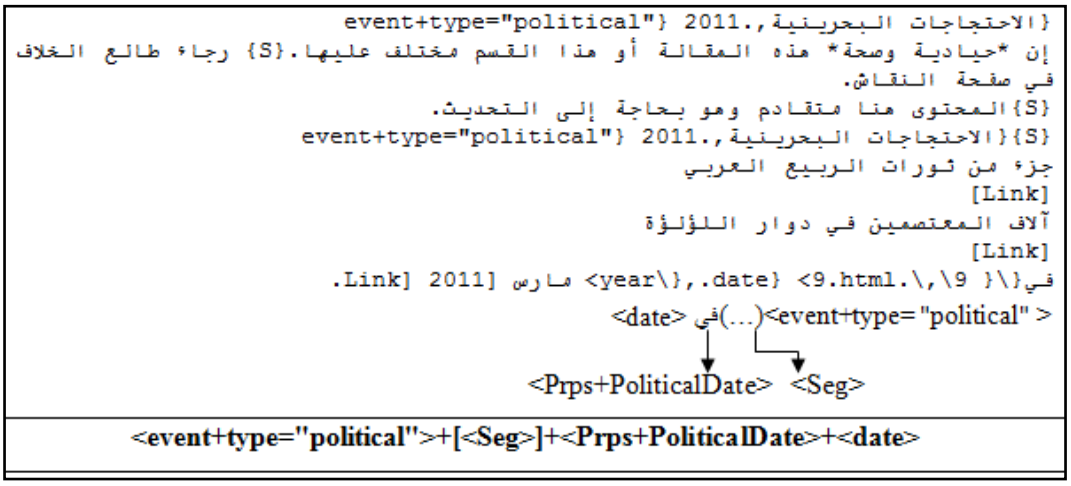


Figure 30. Expression régulière de la RS date politique

La figure 30 montre que nous détectons ce type de RS à travers les extrémités de ce segment sous forme de deux ENA ayant les catégories évènement politique et date. De plus, nous identifions une catégorie lexicale préposition introduisant la date politique. Pour cette raison, nous l'ajoutons comme trait sémantique à l'entrée et nous la stockons dans le dictionnaire sémantique. Cette entrée aura donc la forme suivante : « في/ Prps+PoliticalDate ».

2.3. Expression régulière associée à la RS Familiale

La relation Familiale est une RS exprimant les relations familiales telles que les parents, les enfants, le mariage et le divorce. En fait, cette RS est fréquente dans les articles des noms de personne et est exprimée par des traits sémantiques.

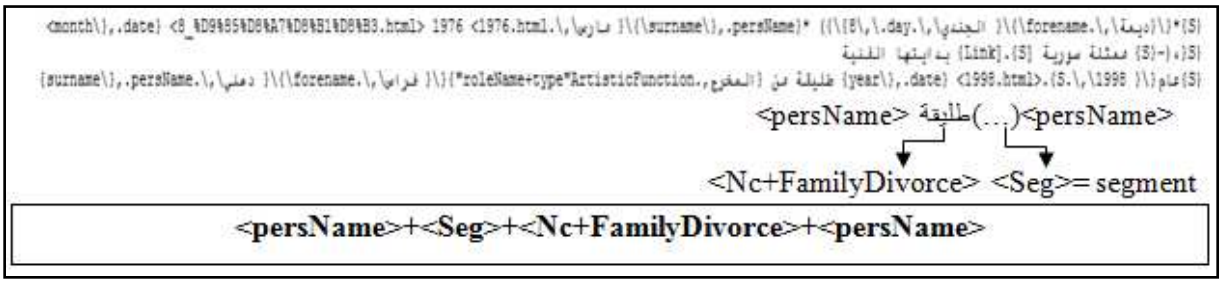


Figure 31. Expression régulière de la RS familiale

La figure 31 montre que nous détectons ce type de relation à travers les extrémités de ce segment dont les deux ont la même catégorie *nom de personne*. De plus, nous détectons le nom commun « طليقة/divorcée » qui joue le rôle d'un mot déclencheur reliant les deux ENA par le sous-type de la RS familiale, ce qui exprime une forte granularité la RS décrite possédant des sous-types tel que divorce, mariage, parents, enfants.

2.4. Expression régulière associée à la RS place politique

Le type de RS place politique permet de décrire un lieu significatif dans lequel un événement politique se déroulait. Dans certains articles décrivant les évènements politiques, nous détectons une structure spécifique qui décrit ce type de RS. Au total, nous avons identifié deux formes d'expressions régulières. Nous expliquons une expression régulière parmi les deux établies à travers l'analyse d'un segment pertinent.

Figure 32. Expression régulière pour la RS place politique

La figure 32 décrit une expression de la RS place politique reliant un événement à un nom de pays. Ces informations sémantiques appartiennent à la partie Info-box d'un article Wikipédia arabe. Pour créer cette expression, nous nous basons sur un nom commun pour détecter la RS du type place politique. Ce nom commun aura le trait sémantique <Nc+PoliticalPlace> pour associer l'évènement à son lieu.

2.5. Exemples d'expressions régulières pour les types de RS identifiées

Nous avons illustré le principe de création des expressions régulières pour quelques types de RS simples. Cependant, l'établissement de ces expressions n'est pas une tâche facile par ce qu'il nécessite un traitement de chaque composant des segments pertinents qui les relient. Dans le tableau suivant, nous choisissons un exemple concret parmi les expressions régulières pour les dix-huit types de RS identifiées.

Tableau 9. Exemples d'expression régulière pour les types de RS identifiées

Relation sémantique	**Exemple d'expression régulière**
Synonymie	<PNC>+<event+type="religious">+<PNC>+<V+Synonymy>+[<Seg>]+ <event+type="religious">+<Prps>+<PNC>+event+type="religious">+ <PNC>
Accessibilité	<placeName+type="mountain">+[Link]+<Nc+Accessibility>+ <placeName+type="city">
Appartenance	<placeName+type="mountain">+[Link]+<PNC>+<Nc+MemberShip>+ <placeName+type="mountain">
Equivalence date	<date>+<Nc+EquivalenceDate>+<date>
Famille	<persName>+<Seg>+<Nc+FamilyDivorce>+<persName>
Date de naissance	{S}+<persName>+<Seg>+<Nc+Birth>+[<PNC>]+<date>
Date de décès	<persName>+<Seg>+<Nc+Death>+<date>
Date politique	<event+type="political">+[<Seg>]+<Prps+PoliticalDate>+<date>
Place politique	<event+type="political">+<Seg>+<Nc+PoliticalPlace> +<placeName+type="country">
Année de fondation	<event+type="cultural">+[<Seg>]+<Nc+YearFoundation>+<date>
Origine	<persName>+[<Seg>]+<V+Origin>+[<Seg>]+<placeName+type="city">
Fonctionnelle	<event+type="cultural">+[<Seg>]+<Nc+Functional>+ [<Seg>]+<persName>
Méronymie	<event+type="cultural">+<Seg>+<Nc+Meronomy>+ <placeName+type="delegation>
Profession	<persName>+[<Seg]+<Adj+Profession>
Nationalité	<persName>+[<Seg>]+<Adj+Nisba>
Age de personne	<persName>+[<Seg>]+<Nc+Age>+[<Seg>]+<NB>+[<Nc+Age>]
Religion	<persName>+[<Seg>]+<Adj+Nisba>
Proximité	<placeName+type="delegation">+[Link]+{S}+{S}+ <placeName+type="country">

Le tableau 9 décrit des expressions régulières contenant divers éléments. Parmi ces éléments, nous trouvons différentes catégories d'ENA comme *nom de personne, évènement*, etc. Nous constatons aussi que ces catégories se situent dans l'extrémité des expressions. De plus, nous observons l'existence des balises de prétraitement comme [Link] et {S}. De plus, nous trouvons des traits sémantiques associés au tagset exploité (relié à Unitex). Il existe aussi les traits sémantiques que nous fixons pour décrire l'information sémantique reliant les ENA. Dans ce

qui suit, nous présentons le dictionnaire sémantique que nous devons créer et nous expliquons les traits appropriés.

3. Modélisation d'un dictionnaire sémantique

La modélisation de notre dictionnaire sémantique est une étape importante permettant de regrouper les indicateurs sémantiques déjà détectés avec des traits significatifs. Ces traits seront appelés ultérieurement lors du processus d'extraction des RS. A travers l'étude linguistique nous avons collecté 116 indicateurs. Nous analysons chaque indicateur rencontré pour dégager le trait grammatical et sémantique convenable. Après la phase d'analyse, nous fixons une syntaxe d'annotation de chaque entrée à stocker dans le dictionnaire sémantique qui respecte la forme suivante : **Entrée de dictionnaire,. Trait grammatical + nom de la RS**

Il faut mentionner que le nom de la RS dans cette syntaxe joue le rôle du trait sémantique pour distinguer les entrées ayant des traits grammaticaux similaires. Dans le tableau suivant, nous regroupons les traits que nous avons modélisés avec leurs significations.

Tableau 10. Traits associés au dictionnaire sémantique

Trait	Signification
Nc/Rltf+MemberShip	Nom commun ou Pronom relatif exprimant l'appartenance
Adj/Nc+Proximity	Nom commun ou Adjectif exprimant la proximité
Prps/Nc/V+Synonymy	Préposition, Nom commun ou Verbe exprimant la synonymie
Nc+FamilyChild	Nom commun exprimant le sous-type enfants de la RS familiale
V/Nc+Origin	Verbe ou Nom commun exprimant l'origine
Nc/V/Prps+Accessibility	Nom commun, Verbe ou Préposition exprimant l'accessibilité
Nc+PoliticalDate	Nom commun exprimant la date politique
Nc/V+Functional	Nom commun ou Verbe exprimant la RS fonctionnelle
Nc/Adj+Profession	Nom commun ou adjectif exprimant la profession
Nc+FamilyMariage	Nom commun exprimant le sous-type mariage de la RS familiale
Nc+Age	Nom commun exprimant l'âge
Nc+PoliticalPlace	Nom commun exprimant la place politique
Nc+EquivalenceDate	Nom commun exprimant l'équivalence des dates
Nc+Meronymy	Nom commun exprimant la méronymie
Nc+FamilyParents	Nom commun exprimant le sous-type parents de la RS familiale
Nc+Birth	Nom commun exprimant l'année de naissance
V/Nc+YearFoundation	Verbe ou Nom commun exprimant l'année de fondation
Adj+Religion	Adjectif exprimant la religion
Nc+Death	Nom commun exprimant la date de décès
Prps+FamilyDivorce	Préposition exprimant le sous-type divorce de la RS familiale

Le tableau 10 décrit les traits que nous avons établis pour décrire une entrée dans le dictionnaire sémantique. La première partie du trait proposé respecte l'annotation associée à la plateforme exploitée. Cependant, la deuxième partie du trait est liée à notre préférence et nous choisissons d'écrire le nom de la RS comme une indication sémantique. Ce dictionnaire sémantique doit être passé en priorité car il peut contenir des entrées existantes dans d'autres dictionnaires exploités pour créer les transducteurs. Par exemple, l'entrée « الأبناء/ les enfants » possède le trait « Nc » dans le dictionnaire des noms communs et « Nc+FamilyChild » dans le dictionnaire sémantique. Dans ce cas, si nous passons le dictionnaire des noms communs en premier, le transducteur dédié à reconnaître la RS familiale ne génère aucun résultat car pour lui le mot « الأبناء » doit avoir l'étiquette « Nc+FamilyChild ». Par conséquent, nous résolvons cette confusion au niveau du nom de chaque dictionnaire. Autrement dit, nous ajoutons le signe « - » à notre dictionnaire sémantique pour qu'il soit renommée comme suit : « DicSemanticPrioritaire-.dic ».

Les traits associés à chaque entrée du dictionnaire sémantique créé jouent le rôle d'un indicateur sémantique aidant les transducteurs à réaliser l'extraction des RS. De plus, ces transducteurs exploitent également les autres traits existants dans les dictionnaires établis et enrichis durant le processus de REN. La REN effectuée sur le corpus d'étude nous offre aussi la possibilité d'utiliser les catégories et les sous-catégories déjà reconnues et annotées. En outre, si un chemin d'extraction rencontre un nom de personne alors il suffit d'écrire dans le nœud le code <persName>. Il faut rappeler dans ce cas que la dernière balise est associée à l'annotation TEI d'un nom de personne. Dans ce que suit, nous passons à la description de la phase d'établissement des transducteurs d'extraction des RS entre les ENA.

4. Etablissement des transducteurs d'analyse

Les transducteurs d'analyse consistent à traduire les expressions régulières à des graphes pour extraire et annoter les RS. Chaque transducteur contient les chemins alternatifs d'extraction en utilisant différents types de nœuds. Ces derniers contiennent soit des traits de dictionnaire soit des variables ou un contenu spécifique indiquant un traitement itératif comme le contexte négatif. Les variables sont utilisées pour organiser l'annotation de sortie à la fin de chaque chemin d'extraction. Le principe de regroupement des chemins ressemble à celui de la REN (proposé dans le chapitre précédant) : les chemins ayant le même parcours sont regroupés ensembles et ceux qui génèrent des ambiguïtés sont séparés.

4.1. Création des sous-graphes

Etant donné que nous avons identifié dix-huit types de RS, nous avons créé 18 transducteurs principaux. Chaque transducteur est responsable de l'extraction et de l'annotation un type

spécifique d'une RS en faisant appel à des sous-graphes qui sont également des transducteurs. En fait, chaque sous-graphe appelé peut contenir à son tour d'autres sous-graphes. Cela dépend de l'imbrication des chemins. Cette répartition en sous-graphes facilite la réutilisation des graphes déjà créés et évite d'avoir des chemins redondants. Passons maintenant à la présentation et à l'explication les transducteurs d'analyse que nous avons établis.

Commençons par le transducteur qui reconnaît la synonymie. Ce transducteur regroupe au total six sous-graphes traitant la Synonymie entre différentes catégories et sous-catégories. Les catégories traitées sont *: nom de ville, nom de personne, évènement, nom de musée, nom de montagne* et *nom de région*.

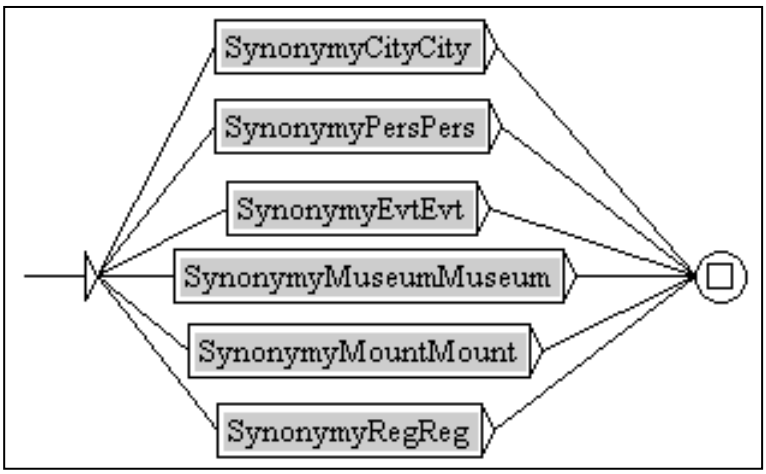


Figure 33. Transducteur principal reconnaissant la synonymie

La figure 33 décrit un transducteur reconnaissant la relation de synonymie. Ce transducteur contient six sous-graphes traitant deux ENA synonymes pour différentes catégories et sous-catégories. Chaque sous-graphe existant dans cette figure se compose d'autres transducteurs.

Nous avons organisé nos transducteurs selon le niveau de raffinement présenté dans notre corpus d'étude. Par la suite, nous illustrons le sous-graphe traitant la synonymie entre les ENA ayant la sous-catégorie nom de ville.

4.2. Exploitation des variables

En se basant sur la notion des variables, nous avons créé les sous-graphes appelés par les graphes principaux dont l'objectif est d'organiser l'annotation de sortie. Dans ce contexte, nous présentons le premier sous-graphe de la figure précédente. Ce sous-graphe ne contient pas d'autres sous-graphes car il existe deux chemins d'extraction de ce type de RS valable pour tous les articles Wikipédia traités. Dans ce cas, nous pouvons dire que le sous-graphe de synonymie entre deux noms de ville admet un aspect générique et applicable indépendamment au domaine d'un texte traité.

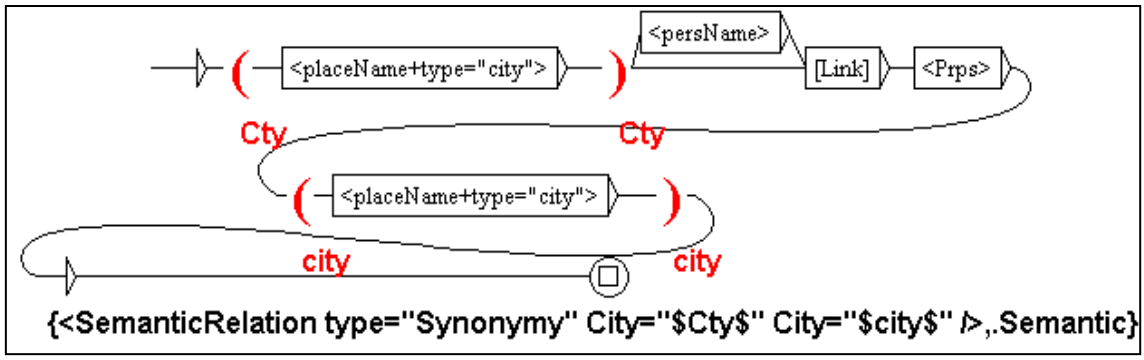


Figure 34. Transducteur reconnaissant la synonymie entre deux noms de ville

La figure 34 décrit le transducteur qui fait l'extraction de la synonymie entre deux ENA ayant la sous-catégorie nom de lieu absolu. Rappelons que le transducteur commence par un nœud qui désigne son état initial. Cet état est suivi d'une boîte contenant l'entrée <placeName+type="city"> située entre les deux symboles « (» et «) » pour dire que cette entrée est déclarée comme une variable. L'avant dernière boîte contient aussi la même valeur que la deuxième et elle est entourée également par une variable décrivant la sous-catégorie nom de ville. Les variables déclarées sont appelées sur le dernier nœud de sortie pour définir la balise d'annotation avec la valeur de chaque ENA et le type de RS qui les relie.

Pour le type de RS méronymie, nous créons un transducteur contenant trois sous-graphes. Le premier graphe extrait les RS entre un évènement et une délégation tandis que le deuxième sous-graphe traite aussi le cas d'un évènement avec une ville. La méronymie entre un évènement et un nom de pays est extraite à travers le troisième sous-graphe. Tous les transducteurs établis utilisent la notion de variable pour organiser l'annotation de sortie. En revanche, il existe d'autres qui profitent de la notion de contexte négatif. Dans ce contexte, nous illustrons un transducteur reconnaissant la méronymie entre un évènement et une ville.

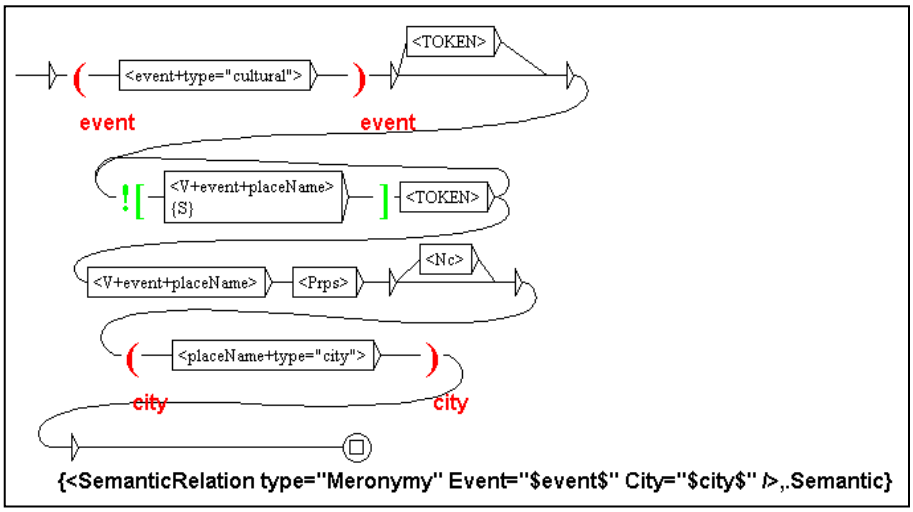


Figure 35. Extraction de la méronymie entre un évènement et une ville

La figure 35 décrit un transducteur reconnaissant la RS méronymie entre deux ENA ayant respectivement la catégorie évènement et la sous-catégorie nom de ville faisant partie des noms de lieu absolu. Ce transducteur contient aussi un chemin d'extraction comportant des variables pour générer une annotation structurée. De plus, les boîtes représentées en couleur verte traduit l'utilisation du concept de contexte négatif à travers les marqueurs **! [**, **]**, conjointement à un chemin qui revient du token sur le contexte. Il s'agit donc d'une forme graphique fonctionnant selon la boucle itérative « tant que ». Ce qui est entouré par les marques **! [**, **]** est la condition pour effectuer le traitement. Dans le transducteur courant, le traitement à faire est la lecture d'un ensemble de tokens (<TOKEN>) tant que le chemin d'extraction n'a pas rencontrée un verbe ayant le trait <V+event+placeName>.

4.3. Exploitation du mode morphologique

Les transducteurs que nous élaborons profitent aussi du principe du mode morphologique. En fait, ce mode morphologique est utilisé souvent pour résoudre un cas de phénomène d'agglutination. Dans l'extraction des RS, l'agglutination se présente au sein du segment pertinent qui relie les ENA. Dans ce qui suit, nous illustrons un exemple d'utilisation du mode morphologique dans un transducteur d'extraction d'accessibilité entre un nom de palais et un nom de délégation.

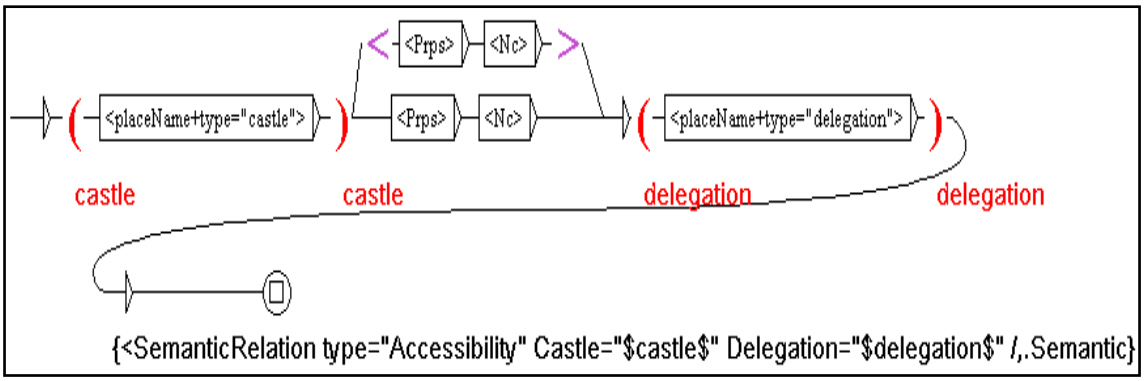


Figure 36. Extraction de l'accessibilité entre les deux noms de lieux

La figure 36 illustre une RS d'accessibilité entre un nom de lieu relatif (nom de palais) et un nom de lieu absolu (nom de délégation). Le mode morphologique dans ce transducteur s'exprime à travers les boîtes contenant les symboles <, >. Ces derniers séparent une préposition agglutinée à un nom commun pouvant indiquer qu'un nom de palais est accessible via un nom de ville.

4.4. Traitement de l'éloignement dans les transducteurs d'analyse

Au cours de l'extraction d'une RS, nous pouvons rencontrer plusieurs problèmes comme l'éloignement de deux ENA qui sont sémantiquement en relation quoiqu'elles soient très

éloignées. Autrement, ces deux ENA se situent sur des segments différents ce qui engendre un débordement de pile au sein des transducteurs. Pour résoudre ce problème, nous proposons la solution suivante : la création d'une balise ayant la forme {chemin d'annotation, Semantic}. Cette nouvelle forme nous permettra également de réduire le temps d'exécution. Dans ce qui suit, nous illustrons un transducteur respectant la nouvelle forme sachant que celle-ci s'écrit au nœud de sortie dédié à l'annotation.

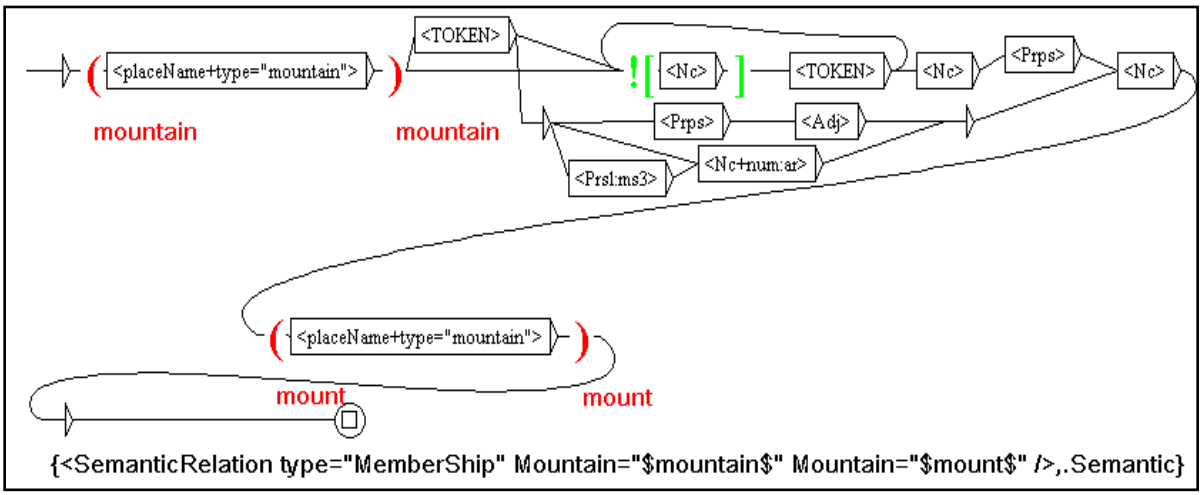


Figure 37. Extraction de la relation d'appartenance entre les noms des montagnes

La figure 37 décrit le transducteur qui extrait la RS appartenance entre deux ENA ayant la même catégorie. Dans tous les transducteurs, nous ajoutons la balise « Semantic » pour protéger l'annotation de sortie ce qui favorise sa réutilisation dans d'autres transducteurs. En général, les transducteurs parcourent le plus long chemin ultérieurement alors l'ordre de passage peut poser le risque aux derniers graphes de ne pas atteindre la bonne destination. Supposons que les dix-sept types de RS se trouvent dans le même texte et qu'elles soient extraites et annotées. Supposons également que la RS numéro dix-huit existe entre une ENA située sur la première ligne en relation avec une deuxième située sur la dernière ligne. Par conséquent, le texte est trop chargé et le transducteur candidat n'arrive pas à les relier pourtant le chemin d'extraction est juste. Pour cette raison, nous exploitons la balise « Semantic ». Elle facilite l'extraction des RS même si les chemins sont très longs.

Afin d'avoir un système harmonieux qui fonctionne convenablement nous devons organiser les transducteurs dans un ordre de passage adéquat. Cet ordre de passage se fixe qu'après plusieurs essais pour garantir un nombre minimal d'erreurs. De plus, l'organisation de ces transducteurs doit faire recours à un mode de passage vu que nos transducteurs couplent les deux processus d'extraction et d'annotation des RS. En outre, ce mode de passage doit fusionner ces deux processus pour aboutir à des RS extraites et annotées au même temps.

5. Importance des RS dans l'enrichissement d'un dictionnaire d'ENA

La recherche des RS aide à réaliser l'enrichissement des dictionnaires électroniques des EN par des détails sémantiques permettant d'élargir leur champ de recherche. Cet enrichissement d'un dictionnaire d'ENA touche son niveau sémantique via l'ajout de nouvelles RS entre les ENA. De plus, ces nouvelles RS peuvent être utilisées pour relier les ENA stockées à des ressources libres externes comme Dbpédia. En outre, une ENA associée à un lien vers une ressource libre peut être recherchée via les RS alternatives qui la relient à une autre ENA. Dans la section courante, nous expliquons l'objectif que nous voulons atteindre à travers un exemple d'enrichissement d'un dictionnaire d'évènements.

Le processus d'enrichissement d'un dictionnaire d'évènements commence par le regroupement d'un ensemble de RS relié à un évènement spécifique. Cet évènement peut être relié à plusieurs catégories comme la catégorie nom de personne. D'ailleurs, les RS sélectionnées ont un type et un sous-type dans certain cas. En fait, cela dépend du type de RS. Par conséquent, nous exploitons deux éléments « type » et « subtype » ainsi leurs valeurs associées pour fournir plus d'informations sémantiques. En effet, une RS se représente comme une balise qui regroupe le type de RS et les catégories reliées.

```
<SemanticRelation type="Functional" sub-type="Director" Event="مهرجان القاهرة السينمائي الدولي" PersName="سمير فريد" />

<SemanticRelation type="Functional" sub-type="Founder" Event="مهرجان القاهرة السينمائي الدولي" orgName="الجمعية المصرية لكتاب ونقاد السينما" />

<SemanticRelation type="YearFoundation" Event="مهرجان القاهرة السينمائي الدولي" Date="1976" />

<SemanticRelation type="Meronymy" Event="مهرجان القاهرة السينمائي الدولي" placeName="القاهرة" />
```

Le segment illustré ci-dessus décrit un ensemble de RS entre deux ENA dont la première entité est toujours de catégorie évènement. A partir de cette figure, nous mentionnons que l'utilisation de différents éléments dans la balise proposée permet de distinguer la nature de la RS ainsi que celle de l'information sémantique extraite.

La deuxième étape dans le processus d'enrichissement consiste à insérer les RS dans les entrées correspondant à un évènement. Dans ce qui suit, nous décrivons un extrait d'une entrée de dictionnaire d'évènements enrichie par des RS.

```
<ANE_entry>
<name>          مهرجان القاهرة السينمائي الدولي </name>
<category> Event </category>
<subCategory>Cultural event </subCategory>
<yearFoundation>1976 </yearFoundation>
<founder>الجمعية المصرية لكتاب ونقاد السينما </founder>
<director>سمير فريد </director>
<placeName>القاهرة </placeName>
</ANE_entry>
```

Le segment précédent montre la richesse en termes de RS entre un évènement et d'autres catégories. Ces catégories liées sont entourées par des balises qui étaient tout d'abord des éléments dans la balise d'identification de la RS. La balise appelée <ANE_entry> décrit une seule entrée et appartient à une balise globale appelé <ANE-list>. Cette dernière possède une DTD qui nécessite une mise à jour après la découverte d'un nouveau type de RS.

La dernière étape dans le processus d'enrichissement est la mise à jour la DTD associée au dictionnaire traité pour mettre à jour les nouveaux éléments ou encore les attributs. Cette mise à jour exige l'ajustement de la structure de dictionnaire puisqu'elle doit subir une phase de validation qui est assurée par des outils liés à la norme d'annotation comme l'outil Oxygen[16] qui valide les fichiers décrits à travers la norme TEI. Cette norme offre également une application pour générer automatiquement les DTD qui s'appelle « Roma [17]».

```
<!ELEMENT ANE_list (ANE_entry+)>
<!ELEMENT ANE_entry
(name,category,subCayegory?,yearFoundation,founder
*, director*,placeName)>
<!ELEMENT name (#PCDATA)>
<!ELEMENT category (#PCDATA)>
<!ELEMENT subCategory (#PCDATA)>
<!ELEMENT yearFoundation (#PCDATA)>
<!ELEMENT founder (#PCDATA)>
<!ELEMENT director (#PCDATA)>
<!ELEMENT placeName (#PCDATA)>
```

La DTD déjà représenté est associé à la structure de dictionnaire stockant les événements. Cette DTD est dédiée à vérifier le dictionnaire d'ENA construit. Ce fichier reflète la catégorisation

[16] https://wiki.tei-c.org/index.php/OXygen
[17] http://www.tei-c.org/Roma/

étendue des ENA vu que nous trouvons deux éléments représentant sa catégorie d'ENA et sa sous-catégorie si elle existe.

6. Conclusion

Dans ce chapitre, nous avons présenté notre démarche proposée pour l'extraction et l'annotation des RS entre les ENA. Cette démarche se base sur la création des ressources nécessaires permettant de construire un ensemble d'expressions régulières et de construire un dictionnaire sémantique prioritaire. Les expressions régulières ont été transformées en des transducteurs d'analyse conçus en se basant sur diverses options. Nous avons établi ces transducteurs en se basant sur la notion des variables pour organiser l'annotation de sortie. De plus, nous avons utilisé la notion de contexte négatif pour éviter certains phénomènes linguistiques. La manipulation de nos transducteurs s'appuyait également sur le traitement de l'éloignement de chemins lors de l'extraction des RS. En outre, nous avons pu résoudre cet éloignement pour extraire le maximum des RS figurant dans un texte arabe donné. Les transducteurs établis ont besoin d'un ordre de passage défini pour qu'ils fournissent les résultats attendus. La génération d'une cascade de transducteurs est nécessaire. Nous avons clos ce chapitre par une explication de l'importance de l'extraction des RS.

L'extraction des RS va nous permettre de maintenir le processus de REN pour corriger les erreurs éventuellement générées. De plus, l'extraction des RS est très utilisée dans diverses applications comme la désambiguïsation. Elle permet également d'indexer les documents qui participent aussi à améliorer la performance des moteurs de recherche. En effet, l'extraction des RS permet d'enrichir la qualité des bases de connaissance, comme les dictionnaires électroniques à travers l'ajout de nouvelles relations pertinentes. En fait, l'enrichissement de ce genre de dictionnaires améliore leur niveau sémantique pour qu'ils fournissent des réponses plus précises. En outre, cet enrichissement par des RS permet de mettre à jour le contenu de ces dictionnaires électroniques d'EN en utilisant des corpus structurés ou semi-structurés. Afin de réaliser notre démarche, nous passons à l'implémentation en nous basant sur la plateforme linguistique Unitex.

Partie 4 : Implémentation, expérimentation et évaluation

Chapitre 7 : Implémentation des systèmes CasANER et ASRextractor

La phase d'implémentation vise à réaliser nos démarches proposées pour obtenir deux systèmes CasANER et ASRextractor dédiés respectivement à la REN et l'extraction des RS entre les ENA. Pour ce faire, nous exploitons la plateforme linguistique Unitex qui offre des techniques avancées pour le traçage graphique des transducteurs déjà conçus. De même, Unitex, nous aide à manipuler les dictionnaires que nous avons modélisés et à améliorer ceux qui existent sous le répertoire Dela associé au module arabe. Quant à l'ordonnancement adéquat des transducteurs, celui-ci est assuré par la notion de cascade qui se génère grâce à l'outil CasSys. Notamment, CasSys applique les transducteurs au sein d'une cascade produite sur les textes selon différents modes de passage. L'acquisition d'un ordre de passage d'une cascade nécessite des tests successifs qui consistent à permuter ses transducteurs jusqu'à atteindre un nombre minimal d'erreurs. L'implémentation touche également des processus de prétraitement et de post-traitement. Le prétraitement se fait à travers une cascade de transducteurs afin de préparer nos corpus pour appliquer après les systèmes élaborés. Le post-traitement comporte un module de normalisation pour transformer les annotations des ENA en TEI. De plus, il inclut un module de récupération des formes brutes des ENA dans des RS extraites via le système ASRextractor.

Le chapitre courant se compose de quatre parties principales dont la première définit les étapes d'implémentation du système CasANER. Dans cette partie, nous présentons la phase de prétraitement effectuée pour préparer le corpus d'étude et du test. Puis, nous décrivons le processus d'enrichissement des dictionnaires existants sous la plateforme Unitex. De plus, nous expliquons la création de nouveaux dictionnaires sachant que ce processus se fait manuellement ou automatiquement en exploitant une banque d'arbres syntaxiques. Nous abordons également l'implémentation du système CasANER qui se compose de cinq modules principaux selon les catégories d'ENA déjà identifiées. Nous passons après à la deuxième partie qui concerne l'implémentation du système ASRextractor qui extrait les RS entre ENA et les annote selon la norme TEI. Dans cette partie, nous décrivons la création d'un dictionnaire sémantique prioritaire permettant de guider le système ASRextractor. Dans la troisième partie, nous présentons la cascade de transducteurs que nous créons pour la normalisation de l'annotation des ENA selon la TEI. Nous clôturons ce chapitre par une phase de post-traitement propre à la représentation des ENA se trouvant au sein des RS extraites par le système ASRextractor.

1. Implémentation du système CasANER

Le système CasANER qui est dédié à la reconnaissance et l'annotation des ENA admet une seule entrée qui est le corpus brut extrait à partir de la Wikipédia arabe. Avant d'effectuer le processus de REN, ce corpus brut passe par une phase de prétraitement visant à le préparer afin d'appliquer le système CasANER. En fait, la sortie du système CasANER est le même corpus

qui contient des ENA reconnues et annotées. Dans la figure 31, nous illustrons l'architecture générale de ce système.

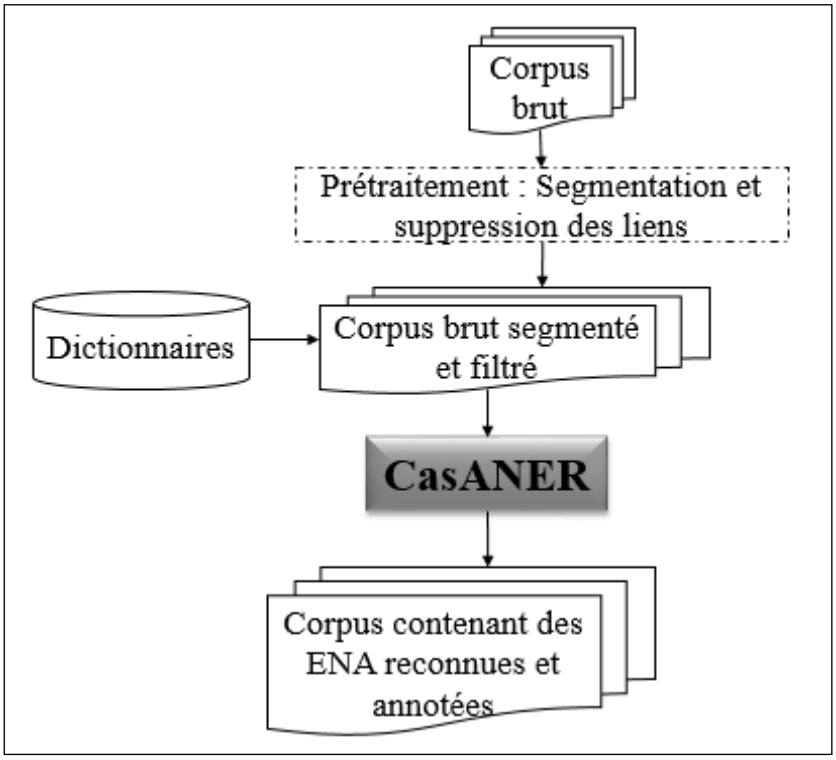


Figure 38. Entrées/sorties du système CasANER

La figure 38 décrit l'entrée du système CasANER qui passe par diverses étapes pour qu'il soit prêt à exploiter. Cette entrée est un corpus brut que nous collectons à partir de la Wikipédia arabe grâce à l'outil Kiwix[18] pour la langue arabe. Il faut mentionner que cet outil nous a aidés à collecter les deux corpus d'étude et de test via une interface facile et manipulable. Avant d'appliquer le système CasANER, le corpus brut doit subir une phase de prétraitement composée de deux sous-phases suivantes : la segmentation et la suppression des liens. Le corpus prétraité nécessite l'application des dictionnaires pour que les tokens qui le composent soient reconnus. Enfin, le corpus résultant du système CasANER va être riche en termes d'ENA reconnues et annotées. Dans ce qui suit, nous expliquons la première partie qui est dédiée à la phase de prétraitement.

2. Phase de prétraitement

La création des corpus (étude et test) à partir de la Wikipédia arabe consiste à collecter des articles provenant de différents pays arabes. Ces articles représentent des pages web mais sous format textuel décrivant plusieurs thématiques qui appartiennent à divers domaines (art, sport, politique, etc.). Nous rappelons que l'acquisition des articles se fait à travers l'outil Kiwix

[18] http://wiki.kiwix.org/wiki/Main_Page/ar

permettant de consulter Wikipédia arabe en mode hors ligne. Via cet outil, nous téléchargeons chaque article sous forme d'un texte ayant l'extension « .txt ». Les articles formant les deux corpus subissent un prétraitement qui est un processus nécessaire pour segmenter les textes en premier lieu et pour éliminer également des liens internes ayant un format spécifique à la structure des articles de la Wikipédia arabe.

2.1. Segmentation du corpus

La segmentation est une étape importante représentant un prétraitement qui peut être effectué sur une ressource textuelle. La segmentation de nos articles sélectionnés se fait grâce à un graphe disponible dans le module arabe de la plateforme linguistique Unitex. Segmenter un texte appartenant à nos corpus consiste à délimiter ses phrases selon les signes de ponctuations. Chaque phrase délimitée prend à sa fin le symbole {S}. Nous illustrons le graphe de segmentation afin de présenter le principe suivi.

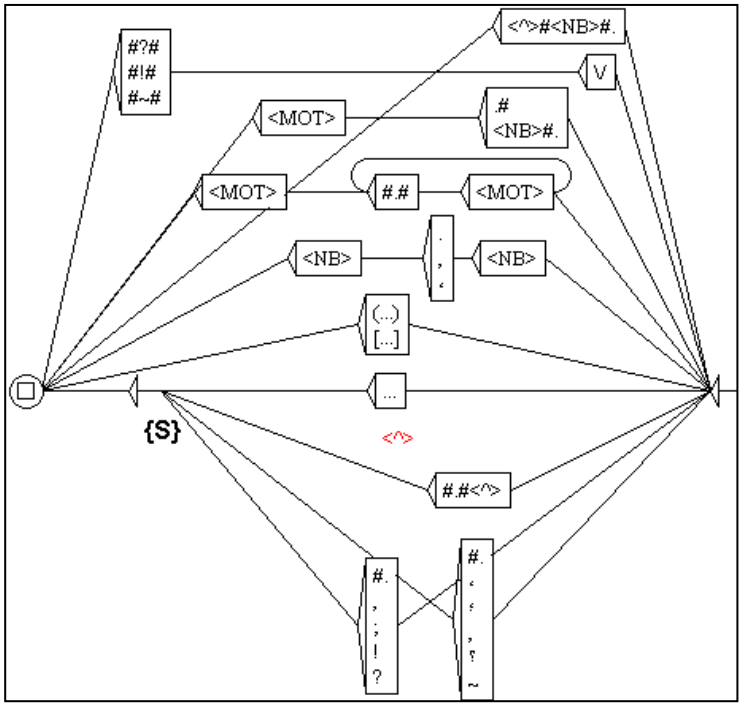


Figure 39. Graphe de segmentation d'un texte arabe intégré sous Unitex

Le graphe de segmentation de la figure 39 est propre à langue arabe. Le graphe implémenté comporte plusieurs chemins. En fait, certains chemins ne possèdent pas une annotation de sortie car ils ne correspondent pas aux signes de segmentation. Parmi les chemins ayant un signe de segmentation, nous constatons l'existence d'une boîte marquée en rouge ayant le contenu « ^ ». Cette boîte signifie le retour à la ligne. Dans notre cas, nous avons décoché ce chemin de reconnaissance vu que nous ne considérons pas le retour à la ligne comme étant la fin d'une

phrase. Rappelons que ce graphe permet de fusionner les séquences reconnues avec l'annotation de sortie définie.

La manipulation et l'application de ce graphe sont assurées par la plateforme Unitex à travers une interface conviviale. La segmentation se propose en chargeant un texte donné sous Unitex à travers la fenêtre « Preprocessing & Lexical parsing » et décrite dans la figure suivante.

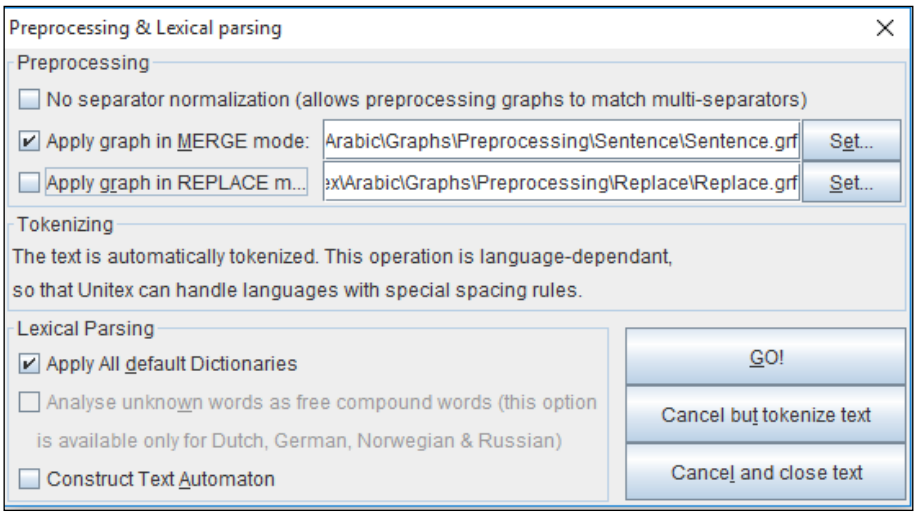


Figure 40. Interface facilitant la segmentation

La figure 40 montre que la segmentation s'effectue en choisissant le graphe adéquat associé à un mode de passage. Deux modes peuvent se présenter soit « merge » pour fusionner le symbole de segmentation et la phrase soit « replace » pour un remplacement. Dans la rubrique « Lexical Parsing », nous pouvons choisir l'application de tous les dictionnaires sur le texte à segmenter.

2.2. Suppression des liens internes

Dans notre travail, le prétraitement ne concerne pas seulement la segmentation des articles mais il englobe également leur filtrage. Nous avons constaté la présence des liens incompressibles et liés à la structure interne de la Wikipédia arabe. Nous avons prédit que ces liens peuvent empêcher le processus de REN surtout que certains liens coupent l'ENA. Pour cette raison, nous avons créé un graphe qui permet de remplacer chaque lien rencontré par un symbole {Link}. Pour mieux expliquer le principe, nous présentons tout d'abord le graphe de la suppression puis nous illustrons un extrait d'un texte.

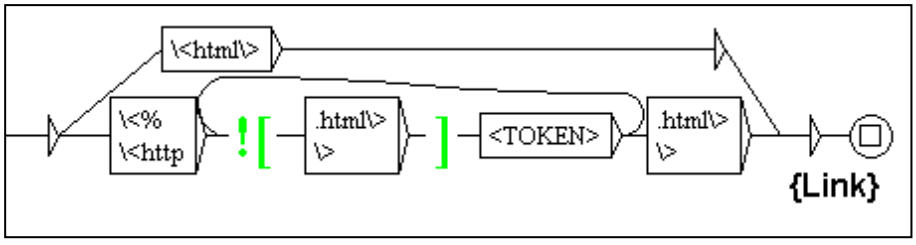


Figure 41. Graphe de suppression des liens internes

La figure 41 montre la forme du graphe qui contient deux chemins possibles. Le premier chemin à parcourir prend la forme de <html> qui doit être remplacé par {Link}. Nous avons créé le deuxième chemin après avoir analysé les différentes formes qu'un lien peut avoir. Ce deuxième commence par la lecture du début du lien comme par exemple « <http ». Puis, nous avons fixé une condition d'arrêt qui à la fin d'un lien candidat pouvant être soit « html> » ou « > ». Par conséquent, chaque token rencontré sera lu tant que la condition d'arrêt n'est pas atteinte. Chaque chemin reconnu par ce graphe illustré sera remplacé par l'annotation de sortie située dans la boîte qui précède le nœud final. Nous donnons un extrait d'un texte montrant le résultat d'application de ce graphe.

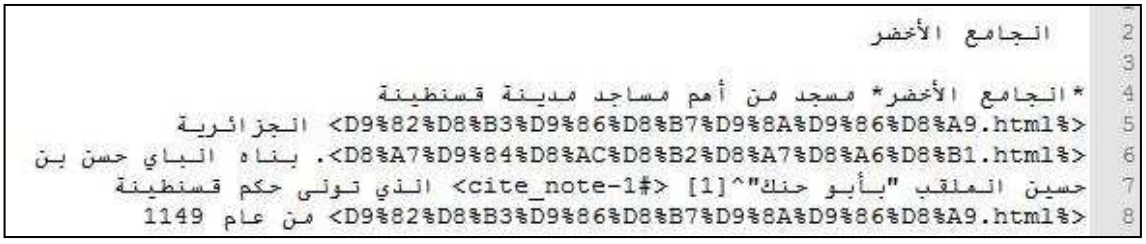


Figure 42. Extrait d'un article de la Wikipédia arabe contenant des liens

La figure 42 illustre le format d'une portion d'un article appartenant à notre corpus d'étude. Nous constatons que les lignes contiennent des liens sous format hexadécimal. Ce format est traité par le graphe de suppression que nous proposons. D'ailleurs, l'application de ce graphe génère le même texte filtré en termes de liens comme elle indique la figure suivante.

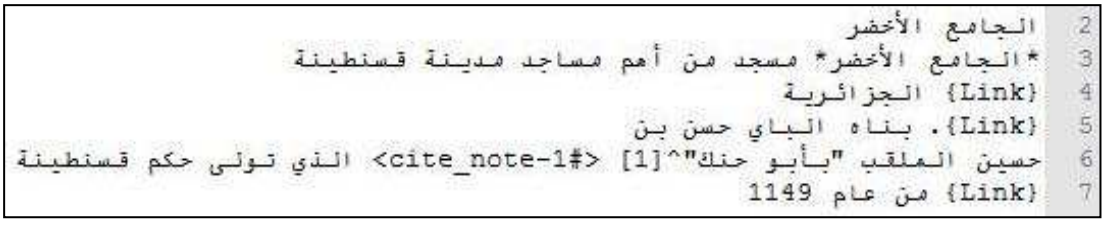


Figure 43. Résultat de la suppression des liens

La figure 43 montre le résultat de la suppression des liens du même extrait illustré précédemment. Les liens figurant dans l'article étaient remplacés par {Link}. En fait, la figure nous permet également de justifier la forme de graphe de la segmentation. Rappelons que nous avons décoché le chemin de retour à la ligne. Si nous observons les lignes 5, 6 et 7 dans la figure courante, nous constatons la présence d'une phrase décomposée en trois lignes.

3. Création et enrichissement des dictionnaires manuellement

La création des dictionnaires a comme objectif de guider le processus de la reconnaissance des ENA, et celui de l'extraction des RS, et de lever certaines ambiguïtés liées aux catégories grammaticales ou encore sémantiques. Rappelons que les dictionnaires sont des fichiers

stockant des entrées ayant des traits qui les catégorisent. Ils possèdent l'extension « .dic » interprétable par Unitex. Les traits décrivant les entrées se divisent en deux types : grammatical et sémantique. D'autres informations additionnelles peuvent s'ajouter également. Les différents traits présentés dans un dictionnaire seront exploités pour la description des constituants d'une règle d'extraction.

3.1. Enrichissement des dictionnaires des noms propres existants

Le processus de REN que nous voulons effectuer nécessite l'existence des dictionnaires des noms propres. Pour cette raison, nous exploitons les dictionnaires créés par [Doumi et al., 2013]. Ces dictionnaires sont accessibles et intégrés sous le répertoire appelé « Dela » sous la plateforme Unitex (Figure 44).

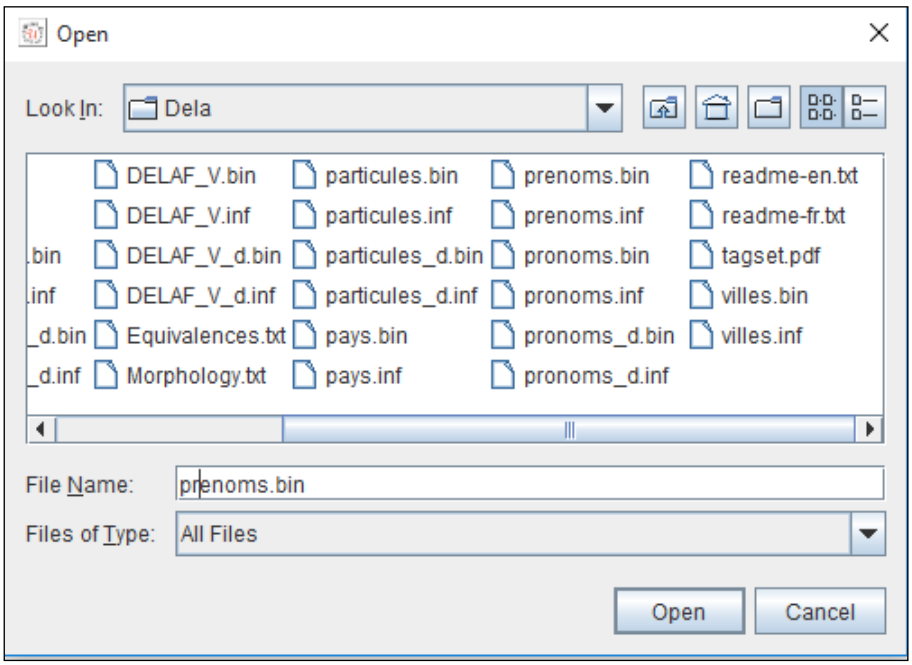


Figure 44. Liste des dictionnaires intégrés sous la plateforme Unitex

La manipulation de ces dictionnaires consiste en une simple utilisation des entrées déjà stockés. Réellement, nous les enrichissons par de nouvelles entrées que nous avons rencontrées dans le corpus d'étude à condition qu'elles n'existent pas dedans. Nous devons mentionner que l'enrichissement de tous les dictionnaires se fait manuellement en ajoutant une nouvelle entrée avec les traits associés. De plus, pour chaque entrée ajoutée nous essayons de traiter ses variantes typographiques. Dans le tableau suivant, nous illustrons la couverture de ces dictionnaires avant et après l'enrichissement.

Tableau 11. Couverture de dictionnaires avant et après l'enrichissement

Dictionnaire	Ancienne couverture	Nouvelle couverture
Prénom	8353	9 927
Nom de lieu	8779	13 760

Le tableau 11 illustre que nous avons pu atteindre une couverture importante qui a été augmentée après des efforts. En fait, nous avons obtenu les anciennes valeurs à partir d'un fichier appelé « readme-fr.txt » qui existe aussi sous le répertoire « Dela ». Pour les noms de lieux, il existe deux dictionnaires séparés pour les noms de pays et ceux de ville. Nous les avons fusionnés pour avoir un seul dictionnaire appelé toponyme. Aux anciennes entrées, nous avons rajouté les noms de délégation et de région.

3.2. Création des nouveaux dictionnaires

Pour enrichir le module arabe de la plateforme Unitex, nous avons créé manuellement de nouveaux dictionnaires. Ces dictionnaires contiennent des entrées, que nous avons collectées lors de l'analyse et de l'exploration de notre corpus d'étude, qui n'existaient pas d'avance. Dans le tableau suivant nous illustrons la couverture de chaque dictionnaire.

Tableau 12. Couverture de dictionnaires créés manuellement

Dictionnaire	Couverture
Nom de famille	1 991
Nom de lieu	13 757
Direction	14
Saison	11
Jour	23
Mois	48

Le tableau 12 décrit la couverture de nos nouveaux dictionnaires qui est importante pour les deux premiers dictionnaires. Les derniers dictionnaires ont une couverture ordinaire vu que le nombre d'entrées est déjà réduit. Nous avons créé automatiquement Adjectif et nom commun et leur mise à jour était manuelle. Dans la section suivante, nous expliquons la création automatique de ces dictionnaires à partir d'une ressource annotée syntaxiquement.

3.3. Création des nouveaux dictionnaires automatiquement

Durant le processus de REN, nous analysons des syntagmes nominaux, adjectivaux et prépositionnels. Parfois, ces syntagmes ne sont pas reconnus et analysés vu que nous envisageons un manque des noms communs et des adjectifs. Pour cette raison, nous exploitons une banque d'arbres syntaxiques appelée ATB (Arabic Tree Bank) pour créer deux dictionnaires stockant respectivement les noms communs et les adjectifs arabes. La création de ces deux dictionnaires s'effectue automatiquement à travers deux extracteurs.

3.3.1. Implémentation d'un extracteur

La liste de noms communs ou des adjectifs générée à partir de l'ATB subit des traitements pour qu'elle forme un dictionnaire exploitable sous Unitex. Dans la figure suivante, nous décrivons les étapes suivies pour atteindre un dictionnaire complet et utilisable.

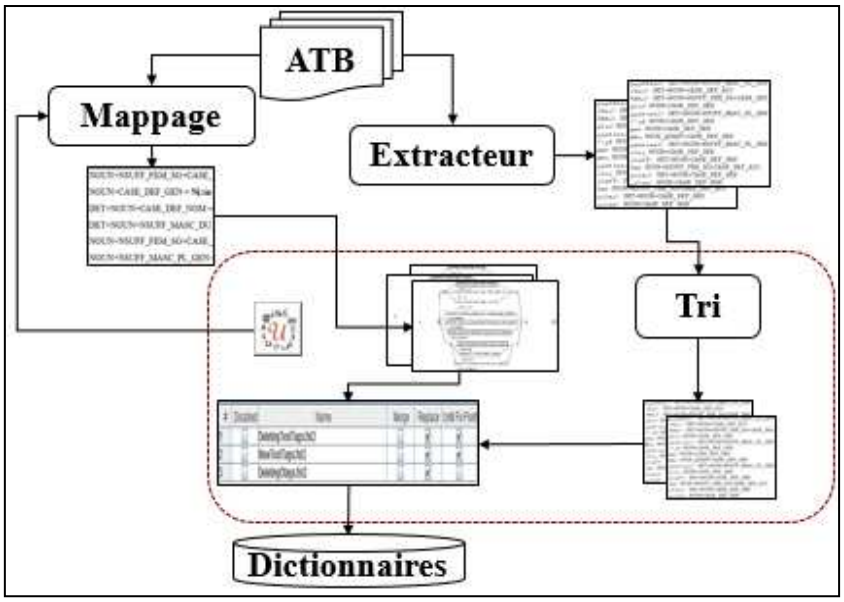


Figure 45. Processus de la création automatique de dictionnaires

La figure 45 décrit les différentes étapes pour créer automatiquement un dictionnaire à partir de l'ATB. La première étape est la génération via un extracteur (programme en Java) de deux fichiers textes à partir de l'ATB. Ces fichiers textes stockent tous les mots ayant la classe grammaticales nom commun ou adjectif ainsi les caractéristiques associées. La deuxième étape est dédiée à filtrer les listes résultantes pour éviter la redondance. Pour cette raison, nous traitons chacune sous la plateforme que nous exploitons vu qu'elle offre cette option. Ce tri nous permet d'avoir deux nouvelles listes contenant deux lignes non redondantes. Ces nouvelles listes sont l'entrée d'un processus d'annotation. En fait, l'annotation des adjectifs se fait manuellement en rajoutant le trait adéquat. Par contre, l'annotation des noms communs constituent la troisième étape illustrée dans cette figure. Cette annotation est assurée par une cascade de transducteurs visant à générer un dictionnaire capable d'être intégré sous la plateforme utilisée. Le principe de l'annotation des noms communs consiste à remplacer le trait associé à l'ATB par celui issu d'une étape de mise en correspondance.

L'extracteur de noms communs que nous avons implémenté est sous la forme d'un programme en java. Cet extracteur prend en entrées les 501 fichiers textes de l'ATB et génère en sortie un seul fichier contenant que les noms communs et traits associés.

```
public static void main(String[] args){
    File directory_in=create(new File("E:\\doctorat\\documents_copies_de_mon_flash\\atb_2_3.1\\data\\pos\\after"));

    File[] fichiers=directory_in.listFiles();
    //System.out.println("path file : "+this.getAbsolutePath());
    String path_out=directory_in.getParent()+"\\Noms_communs_atb.txt";
    File file_out=new File(path_out);
    try {
        BufferedWriter buff_out = new BufferedWriter(new FileWriter(path_out));
    for(int x=0;x<fichiers.length;x++){
        try {
            BufferedReader buff_in = new BufferedReader(new FileReader(fichiers[x].getAbsoluteFile()));
            String line, inputstring1="", inputstring2;
            while ((line = buff_in.readLine()) != null) {
                if(line.startsWith(" INPUT STRING: ")){
                    inputstring1=line.replace(" INPUT STRING: ", "");
                }
                if(line.startsWith("          POS: ")){
                    inputstring2=line.replace("          POS: ", "");
                    if(inputstring2.contains("NOUN") && !inputstring2.contains("NOUN_PROP")){
                        buff_out.write(inputstring1+" "+inputstring2);
                        if((line = buff_in.readLine()) != null)
                            buff_out.newLine();
                    }
                }
            }
```

Figure 46. Extrait du code de l'extracteur de noms communs

L'exécution de l'extracteur illustré dans la figure 46 que nous avons implémenté a donné une liste contenant au total 60 045 noms communs. Les noms communs extraits peuvent se répéter dans plusieurs textes de l'ATB. Pour cette raison, nous avons besoin d'un processus de tri pour supprimer les noms communs redondants.

3.3.2. Processus de tri des noms communs issus de l'ATB

Le tri des fichiers résultants de l'extracteur à travers la plateforme linguistique Unitex. La nouvelle liste générée contient 13 958 noms communs. Nous constatons alors qu'il existait 46 087 lignes redondantes. La liste n'est pas encore filtrée convenablement. En fait, il faut appliquer les dictionnaires des noms communs (existant sous Unitex) sur cette liste afin de détecter pour garder les entrées trouvées vu qu'elles n'existent pas dans les dictionnaires appliqués. C'est un traitement simple consistant à charger le texte sous Unitex en suivant les étapes de prétraitement (quoique sans appliquer les graphes de segmentations). Dans le dossier de prétraitement (snt), le fichier « err » est le fichier que nous voulons récupérer son contenu. Les entrées non reconnues sont 5 116 entrées.

3.3.3. Mise en correspondance des traits de l'ATB et Unitex

L'annotation des entrées représentant les noms communs est encore propre à l'ATB. Pour cette raison, une phase de mise en correspondance entre les traits de l'ATB et ceux du tagset[19] que nous exploitons est nécessaire. Cette mise en correspondance nous a permis de dégager différentes règles d'extraction. Ces règles d'extraction sont réparties en deux ensembles : des règles de transformation, au total 95 règles, et des règles de filtrage, en total 7 règles.

Tableau 13. Extrait de processus de mappage

Tait associé à l'ATB	**Nouveau trait associé à Unitex**	**Exemple**
NOUN+CASE_DEF_NOM	Nc:ur	بنك
DET+NOUN+CASE_DEF_GEN	Nc:ir	المصارف
NOUN+CASE_DEF_ACC	Nc:an	مجلس
NOUN+NSUFF_FEM_SG+CASE_DEF_GEN	Nc:fsir	إدارات
NOUN+CASE_DEF_GEN	Nc:in	اتحاد
DET+NOUN+CASE_DEF_NOM	Nc:ur	الدكتور
DET+NOUN+NSUFF_MASC_DU_NOM	Nc:mdr	المصرفان
NOUN+NSUFF_FEM_SG+CASE_INDEFF_GEN	Nc:fsin	نسبة
NOUN+NSUFF_MASC_PL_GEN	Nc:mpin	ممثلي
NOUN_NUM+NSUFF_FEM_SG+CASE_DEF_ACC	Nc:fsan	أربعة
NOUN+CASE_INDEF_GEN	Nc:in	أعضاء
NOUN_NUM+CASE_INDEF_NOM	Nc+num:un	آلاف
NOUN+NSUFF_FEM_PL+CASE_DEF_ACC	Nc:fpar	آليات
NOUN+CASE_INDEFF_ACC	Nc:ar	أبعادًا

Le tableau 13 illustre un extrait de processus de mappage après une analyse approfondie. Ce processus de mappage nous a aidé à créer des règles qui prennent en entrée un trait de l'ATB et le remplacent en sortie par celui respectant le tagset d'Unitex. Les règles exploitées n'admettent pas un trait sémantique sauf celles qui traitent les noms communs numériques comme « آلاف » ou « أربعة ». Le rôle des règles de filtrage que nous créons consiste à éliminer quelques formes de mise en correspondance pouvant engendrer des ambiguïtés. De plus, ces règles peuvent filtrer également des traits déjà transformés et annotés en erreurs.

[19] Le tagset des dictionnaires de module arabe sous Unitex

3.3.4. Création des graphes de transformation et filtrage de traits

Toutes les règles que nous avons établies sont représentées sous forme de graphes ayant deux ensembles d'appartenance : des graphes de transformation et des graphes de filtrage. Les deux types de graphes fonctionnent selon le même principe car ils prennent en entrée les traits associés à l'ATB et génère en sortie les traits souhaités.

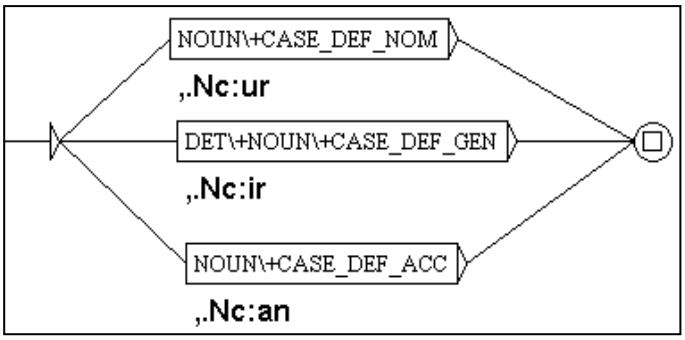


Figure 47. Graphe de transformation trait de l'ATB vers Unitex

La figure 47 illustre un graphe parmi dix graphes permettant de transformer les annotations des noms communs. Les chemins de ce graphe correspondent aux règles d'extraction obtenues après le processus de mappage effectué précédemment.

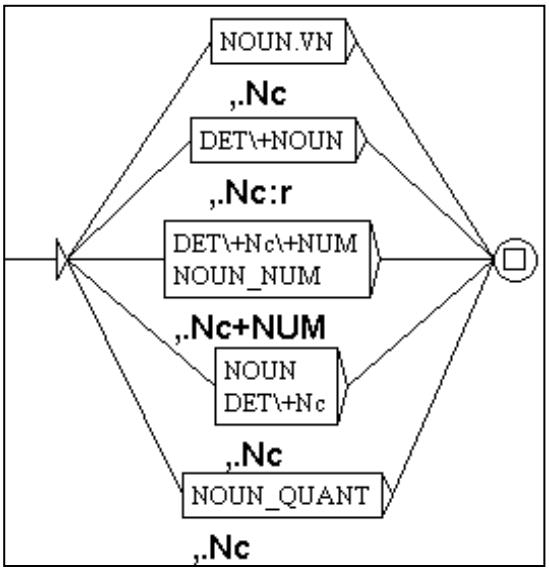


Figure 48. Graphe de filtrage de trait

La figure 48 décrit les sept chemins de filtrage effectués grâce aux règles d'extraction identifiées. Dans certains chemins, il s'agit juste d'un remplacement simple. Ce remplacement est justifié par le fait que le trait concerné n'admet pas de correspondance dans le tagset. Prenons l'exemple du dernier chemin, il prend « NOUN_QAUNT » et génère « ,.Nc+quant ».

3.3.5. Création d'une cascade de transformation de traits

Les transducteurs que nous avons élaborés sont regroupés au sein d'une cascade contenant un graphe principal appelé « AtbToUnitexTransformation » regroupant 10 graphes principaux dédiés à la transformation et un graphe appelé « AtbToUnitexFiltering » pour le filtrage. Tous les graphes regroupés sont passés selon le mode « replace ». Pour montrer le résultat de la création automatique de nos deux dictionnaires, nous choisissons d'illustrer un extrait de la liste des noms communs issu de l'ATB illustré dans la figure 49 et un autre extrait du dictionnaire créé automatiquement dans la Figure 50.

```
البلاد DET+NOUN+CASE_DEF_ACC
الخطة DET+NOUN+NSUFF_FEM_SG+CASE_DEF_NOM
تدخل NOUN+CASE_DEF_GEN
المساهمين DET+NOUN+NSUFF_MASC_PL_GEN
شراء NOUN+CASE_DEF_GEN
حصص NOUN+CASE_DEF_GEN
بعض NOUN_QUANT+CASE_DEF_GEN
المساهمين DET+NOUN+NSUFF_MASC_PL_GEN
رجال NOUN+CASE_DEF_GEN
```

Figure 49. Extrait de la liste des noms communs issu de l'ATB

```
آئمة,.Nc
آثار,.Nc:an
آثار,.Nc:in
آثار,.Nc:ur
آجال,.Nc:in
آداب,.Nc:in
آذان,.Nc:in
آذان,.Nc:un
آراء,.Nc:an
آراء,.Nc:in
آراء,.Nc:un
```

Figure 50. Extrait de la liste issu du processus de création

Après la transformation et le filtrage effectué grâce à notre cascade de transducteurs, la liste qui résulte de la cascade est un dictionnaire qui peut subir des mises à jour régulières.

3.3.6. Création d'un dictionnaire d'adjectifs

La création d'un dictionnaire stockant les adjectifs arabes est plus simple que le dictionnaire de noms communs. En fait, nous avons exploité le même extracteur en changeant la nature de trait à extraire de NOUN (nom commun) à ADJ (adjectif). Dans la figure suivante, nous illustrons

un extrait de liste générée après l'exécution de l'extracteur en java alimenté par le nouveau paramètre (adjectif).

ثاني ADJ_NUM
أكبر ADJ_COMP+CASE_DEF_GEN
العامة DET+ADJ+NSUFF_FEM_SG+CASE_DEF_GEN
المصرية DET+ADJ+NSUFF_FEM_SG+CASE_DEF_GEN
الدولي DET+ADJ+CASE_DEF_ACC
المصرية DET+ADJ+NSUFF_FEM_SG+CASE_DEF_GEN
العليا DET+ADJ
جديد ADJ+CASE_INDEF_GEN
المتعثرة DET+ADJ+NSUFF_FEM_SG+CASE_DEF_GEN
الخطرة DET+ADJ+NSUFF_FEM_SG+CASE_DEF_GEN

Figure 51. Extrait de la liste des adjectifs générée à partir de l'ATB

La figure 51 illustre les onze premières lignes d'une liste d'adjectifs arabes ayant 18 108 lignes. Pour éliminer les lignes redondantes, nous l'avons chargée sous Unitex pour la trier. À la suite de ce tri, il ne nous reste que 5 051 lignes non redondantes. Etant donné qu'Unitex contient un dictionnaire d'adjectifs arabes, nous avons appliqué ce dernier sur la liste trié. Nous sommes allés dans le dossier de prétraitement (snt) de cette liste chargée sous Unitex pour consulter le fichier « err ». Ce fichier va contenir les entrées non reconnues par le dictionnaire d'adjectif arabe existant et effectivement nous avons trouvé que 2 087 lignes non reconnues. Finalement, nous avons rajouté le trait « ,.Adj » à ces nouvelles lignes.

Les dictionnaires que nous avons créés automatiquement ont subi à leur tour des améliorations au niveau de leur couverture. Pour visualiser l'augmentation du nombre d'entrées nous proposons le tableau suivant.

Tableau 14. Couverture des dictionnaires créés automatiquement

Dictionnaire	**Couverture initiale**	**Nouvelle couverture**
Adjectif arabe	2 087	2 938
Nom commun	5 116	14 976

Le tableau 14 montre le résultat de nos efforts fournis pour enrichir les dictionnaires qui avaient une couverture moyenne. Les entrées rajoutées ont été découvertes lors de l'exploration de notre corpus d'étude qui est extrait à partir de la Wikipédia arabe.

4. Implémentation proprement dite

Le système CasANER est une cascade de transducteurs comportant cinq modules principaux. Ces modules sont organisés selon un ordre précis fixé selon plusieurs tests. Chaque module décrit une catégorie principale faisant partie de la typologie d'ENA que nous avons proposé

dans le chapitre trois. Dans la figure suivante, nous décrivons l'architecture de système CasANER et les modules qui le composent.

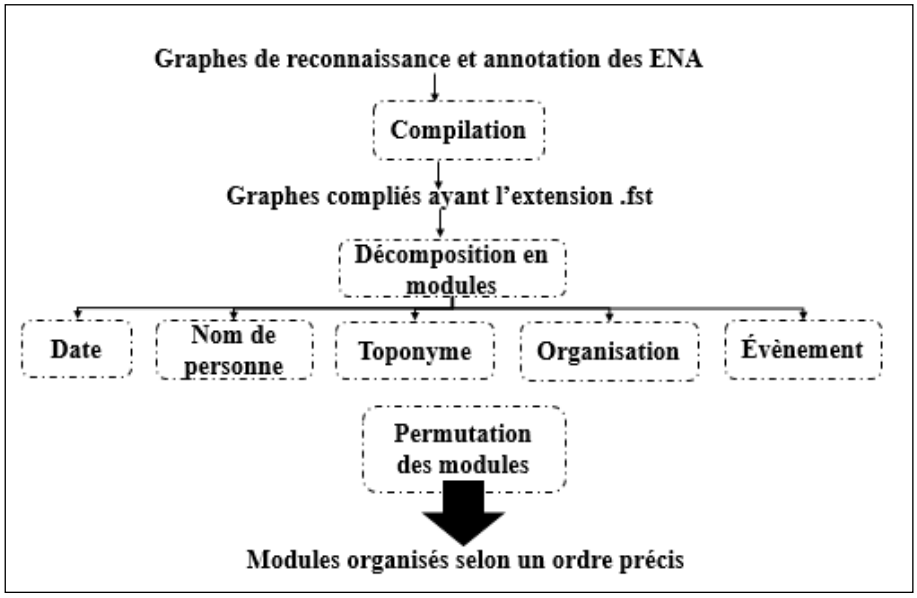


Figure 52. Architecture du système CasANER

La figure 52 montre que les graphes déjà élaborés nécessitent un processus de compilation pour qu'ils soient regroupés dans notre cascade de transducteurs avec l'extension « .fst ». Par conséquent, ces graphes s'organisent sous forme de modules renommés selon les catégories principales d'ENA. Nous notons que chaque module peut être exploité indépendamment des autres modules. Toutefois, le choix de l'ordre de passage est important lorsqu'on applique plus qu'un module. En fait, avoir un ordre de passage précis permet l'optimisation du temps d'exécution en respectant la structure du texte traité. De plus, cet ordre aide à générer un nombre d'erreurs minimales.

Certains modules dans la cascade de transducteurs du système CasANER contiennent les trois ensembles de graphes qui sont : analyse, filtrage et généralisation d'étiquetage. Les graphes d'analyse nécessitent la fusion des étiquettes d'annotation et les ENA dans les textes. Mais, les graphes de filtrage remplacent les anciennes ENA avec celles reconnues et annotées. La cascade de transducteurs, générée grâce à l'outil CasSys, regroupe quinze graphes principaux alors que nous avons élaboré 178 graphes au total. La génération de la version finale de la cascade de transducteurs composant le système CasANER n'est pas une tâche facile. En réalité, nous devons changer l'ordre des graphes jusqu'à fixer un ordre adéquat. Chaque changement subit une phase de test pour voir si c'est le bon choix ou il faut réordonner de nouveau ces graphes.

#	Disabled	Name	Merge	Replace	Until Fix Point	Generic
1	☐	Date.fst2	☑	☐	☐	☐
2	☐	DateFiltering.fst2	☐	☑	☐	☐
3	☐	GenericGraphFordate.fst2	☑	☐	☐	☑
4	☐	PersName.fst2	☑	☐	☐	☐
5	☐	PersNameFiltrage.fst2	☐	☑	☐	☐
6	☐	ToponymAbsolute.fst2	☑	☐	☐	☐
7	☐	HydronymRec.fst2	☑	☐	☐	☐
8	☐	ToponymRelative.fst2	☑	☐	☐	☐
9	☐	ToponymRecPart2.fst2	☑	☐	☐	☐
10	☐	ToponymRelativeFiltering.fst2	☐	☑	☐	☐
11	☐	OrgName.fst2	☑	☐	☐	☐
12	☐	ReligiousEvent.fst2	☑	☐	☐	☐
13	☐	PoliticalEvent.fst2	☑	☐	☐	☐
14	☐	CulturalEvent.fst2	☑	☐	☐	☐
15	☐	GenericGraphForPersName.fst2	☑	☐	☐	☑

Figure 53. Cascade de transducteurs représentant le système CasANER

La figure 53 montre l'ordre de passage que nous avons fixé pour appeler les graphes principaux. En fait, seulement les graphes d'analyses sont passés qu'en mode « merge » qui permet de fusionner les ENA reconnues avec leurs annotations. Cependant, nous avons utilisé le mode « replace » pour appeler les graphes de filtrages. Ce genre de graphes nécessite d'être passé selon ce mode vu qu'ils exploitent des variables pour structurer la sortie et remplacer l'ancienne ENA par une nouvelle représentation correcte et annotée. Les graphes d'étiquetage génériques utilisent le mode « merge » aussi mais ils doivent s'exécuter en mode « Generic » pour respecter leur forme spécifique que nous avons illustrés auparavant.

Durant la phase de test, nous avons détecté plusieurs types d'erreurs. Par exemple, en traitant la catégorie *Evènement* nous devons séparer les graphes selon les sous-catégories (culturel, politique et religieux). En fait, si nous insérons le graphe traitant les évènements culturels avant celui reconnaissant les évènements religieux alors certaines ENA ne seront pas reconnues. Par exemple, l'ENA « مهرجان عيد الفطر/ Festival Eid Al-fitr » ne sera pas reconnue par ce qu'il est composé par un évènement religieux « عيد الفطر / Eid Al-fitr ». L'ordre de passage ne dépend pas seulement des erreurs concrètes générées par cette cascade mais il dépend également d'une certaine logique. En outre, il faut passer les graphes qui seront appelés au sein des autres graphes pour assurer une bonne reconnaissance. Par exemple, nous passons le graphe principal des noms de personnes avant les noms de lieux relatifs car il existe des noms lieux faisant référence à une fameuse personne.

5. Implémentation du système ASRextractor

Le système ASRextractor vise à extraire les RS entre les ENA et les annoter en TEI. Le système ASRextractor prend en entrée un texte annoté selon le système CasANER et génère en sortie le même texte avec des RS extraites et annotées. Le texte d'entrée subit une application d'un

dictionnaire sémantique pour détecter les indicateurs sémantiques aidant les graphes à fonctionner ultérieurement.

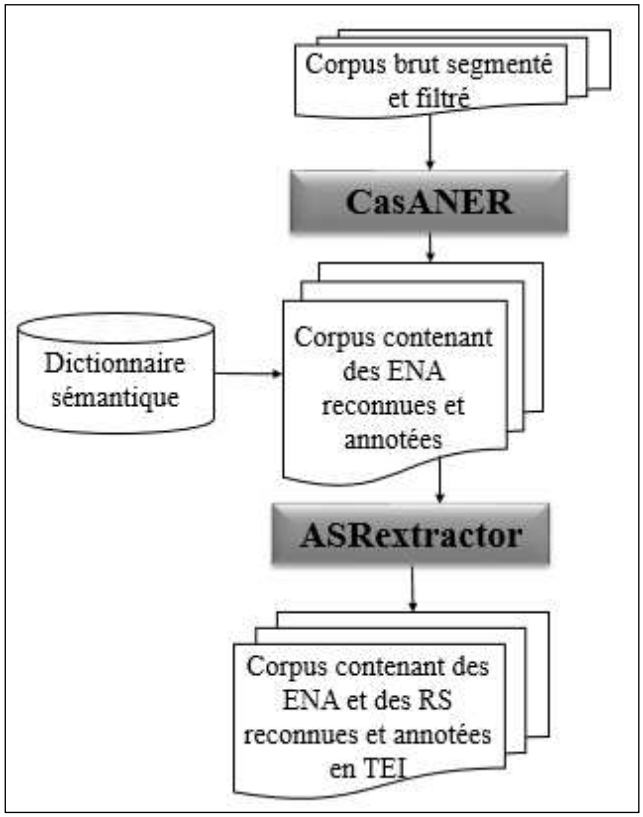


Figure 54. Entrée/Sortie du système ASRextractor

La figure 54 montre que les deux systèmes que nous avons élaborés sont fortement liés vu que la sortie du système CasANER est l'entrée du système ASRextractor. L'entrée de ce dernier système a besoin d'un dictionnaire sémantique pour guider le processus d'extraction des RS. Cependant, les autres dictionnaires sont déjà appliqués d'une façon automatique lors du chargement d'un texte à annoter.

5.1. Création d'un dictionnaire sémantique prioritaire

Sous la plateforme Unitex, nous avons la possibilité de choisir l'ordre de passage des dictionnaires créés sur les textes déjà traités. En fait, nous pouvons utiliser le signe plus « + » qui rend le dictionnaire mentionné établi au dernier ordre. De même, nous pouvons utiliser le signe moins « - » qui rend le dictionnaire mentionné établi au premier ordre. Dans ce contexte, nous allons exploiter le deuxième signe pour le dictionnaire sémantique que nous créons pour guider le processus d'extraction des RS. Notre dictionnaire sémantique a besoin d'étiqueter en premier les tokens qui se présentent dans le texte traité. En outre, ce dictionnaire peut stocker des entrées qui existent déjà dans d'autres dictionnaires mais qui jouent le rôle des indicateurs sémantiques. Dans la figure suivante, nous illustrons la forme de ce dictionnaire sémantique.

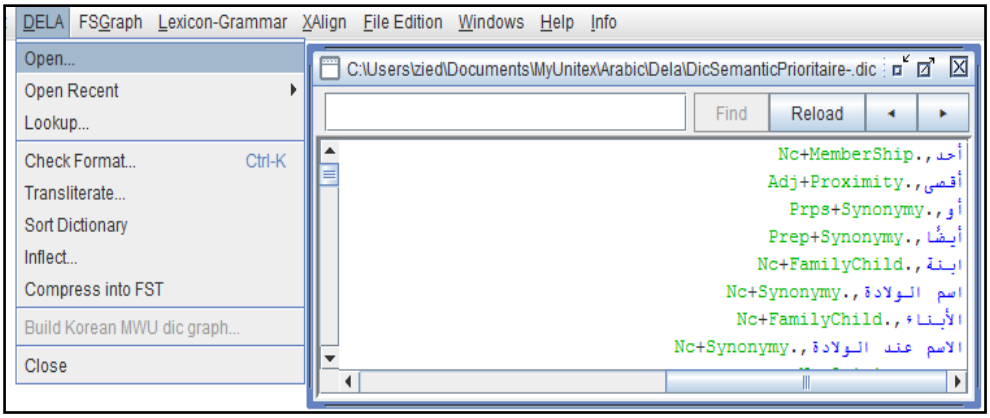


Figure 55. Chargement du dictionnaire sémantique prioritaire

La figure 55 montre un extrait de notre dictionnaire sémantique qui est appelé « DicSemanticPrioritaire- » sachant qu'il stocke 116 entrées. Pour chaque entrée, nous utilisons les descriptions des mots existants dans le tagset en ajoutant la spécification sémantique correspondante. Par exemple, dans notre corpus d'étude la préposition « أو/ou » est utilisée fréquemment pour relier deux ENA ayant la même catégorie. Donc, cette préposition joue le rôle d'un déclencheur de la RS Synonymie. Pour cette raison, elle est définie dans le dictionnaire sémantique par <Prps+Synonymy>. Aussi, le nom commun « أحد/ l'un » existe dans la plupart des articles lié à l'ENA ayant la sous-catégorie *Montagne*, ce qui permet de déduire qu'il est un déclencheur de la RS Appartenance. Alors, nous ajoutons dans notre dictionnaire le trait sémantique <Nc+MemberShip> qui sera exploité par les transducteurs définissant la RS.

5.2. Implémentation du système ASRextractor

Le système ASRextractor vise à extraire et annoter dix-huit types de RS entre les ENA. Pour cette raison, ce système se compose de 18 modules qui sont indépendants. En outre, chaque module peut être exploité selon le type à extraire. Dans notre cas, les 18 modules sont organisés selon un ordre choisi à la base de la structure d'un article Wikipédia et selon sa nature également. De plus, cet ordre a subi divers essais pour permuter quelques modules jusqu'à atteindre une sortie ayant un taux d'erreurs réduit. Nous rappelons que le système ASRextractor est composé d'une cascade de transducteurs regroupant 18 graphes principaux. Par contre, nous avons élaboré 199 graphes au total. Dans la figure suivante, nous illustrons la forme de cette cascade de transducteurs d'analyse.

#	Disabled	Name	Merge
1	☐	PersAge.fst2	☑
2	☐	DeathPersName.fst2	☑
3	☐	Family.fst2	☑
4	☐	BirthPersName.fst2	☑
5	☐	OriginPersName.fst2	☑
6	☐	YearFoundation.fst2	☑
7	☐	MemberShip.fst2	☑
8	☐	Synonymy.fst2	☑
9	☐	Profession.fst2	☑
10	☐	Nationality.fst2	☑
11	☐	Religion.fst2	☑
12	☐	Fonctional.fst2	☑
13	☐	Accessibility.fst2	☑
14	☐	Meronymy.fst2	☑
15	☐	EquivalenceDate.fst2	☑
16	☐	Proximity.fst2	☑
17	☐	PoliticalDate.fst2	☑
18	☐	PoliticalPlace.fst2	☑

Figure 56. Cascade de transducteurs représentant le système ASRextractor

La figure 56 montre la forme de la cascade de transducteurs composant le système ASRextractor qui est dédié à l'extraction des RS entre les ENA. ASRextractor permet d'extraire dix-huit types de RS à base des graphes ordonnés avec un ordre bien précis. Chaque graphe dans la cascade ajoute ses propres annotations sur le texte courant selon le mode de passage « Merge ». Ce mode délivre en sortie une RS reconnue et définie par une balise TEI.

6. Implémentation de la cascade de normalisation des ENA

La cascade de normalisation permet de transformer l'annotation des ENA reconnues en utilisant la norme TEI. Cette transformation permet de générer une sortie structurée en remplaçant les balises associées à l'outil CasSys par celles de la TEI. Comme les autres cascades déjà mentionnées, la cascade de normalisation nécessite un ordre de passage. Par contre, le choix de cet ordre n'est pas compliqué car il n'existe que deux graphes dont leurs rôles exigent l'ordre adéquat. Notamment, le mode passage est bien choisi vu que les graphes de synthèses utilisent la notion des variables et traite aussi les imbrications au sein d'une ENA. Dans la figure suivante, nous illustrons la forme de cette cascade.

#	Disabled	Name	Merge	Replace	Until Fix Point	Generic
1	☐	balisageType.fst2	☐	☑	☑	☐
2	☐	balisage.fst2	☐	☑	☑	☐

Figure 57. Cascade de normalisation de l'annotation des ENA en TEI

La figure 57 illustre l'ordre de passages des deux graphes formant la cascade de normalisation. En outre, nous choisissons de passer en premier le graphe reconnaissant les ENA dont leurs annotations ont des éléments. Dans cette cascade, nous utilisons un nouveau mode de passage qui est « Until Fix point » permettant de traiter les occurrences des ENA dans un texte donné.

Le graphe selon ce mode s'applique en itération jusqu'à ne plus avoir de changements au niveau de l'annotation. Pour résumer le rôle de cette cascade, nous proposons la figure suivante présentant un exemple illustratif.

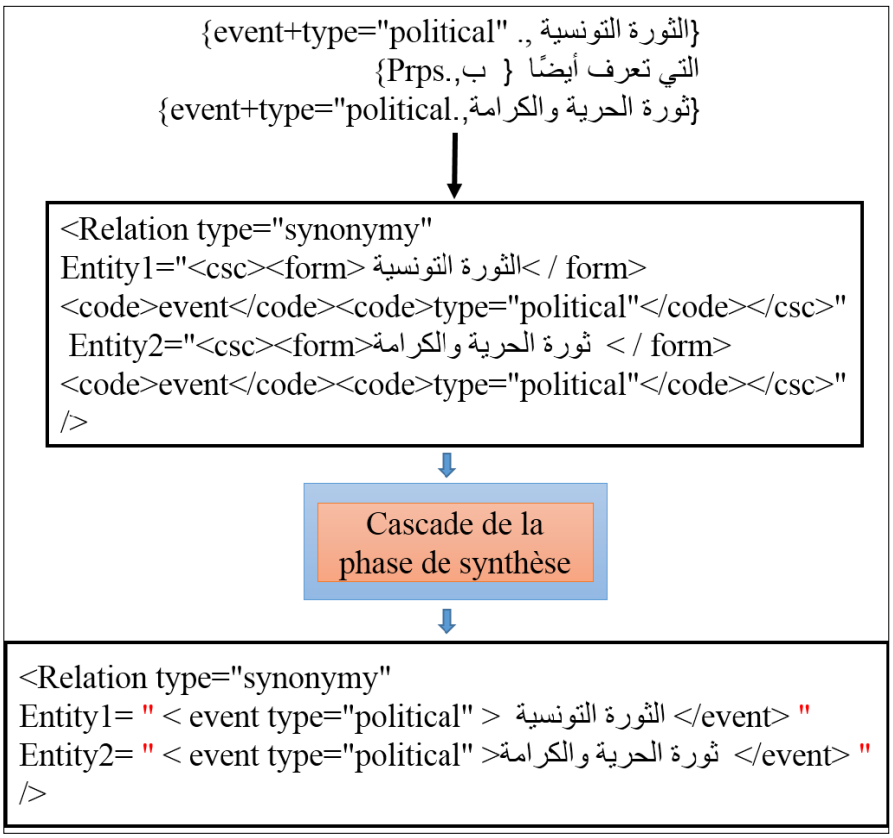


Figure 58. Exemple de normalisation de l'annotation des ENA

La figure 58 montre que l'ENA est initialement annotée avec des accolades {} qui est une représentation spécifique à Unitex. L'annotation via les accolades possède une traduction décrite en XML dont les balises ont un format spécifique défini par CasSys. Par exemple, un fichier, issu d'une application d'une cascade de REN implémenté à travers CasSys, contient des balises XML spécifiques comme <csc> et </csc> entourant l'ENA reconnue via cette cascade. Ces balises peuvent englober d'autres formes de balises comme <form> et </form> pour décrire l'instance d'une ENA et <code> et </code> décrivant la catégorie dans laquelle une ENA est assignée. La dernière balise peut contenir une autre balise ayant la même appellation <code> et </code> pour citer une sous-catégorie de l'ENA. Finalement, la cascade de synthèse va tirer profit de cette représentation pour avoir une représentation structurée selon la TEI.

7. Post-traitement relié aux ENA dans les RS

Le post-traitement à effectuer dans cette sous-section consiste à récupérer la forme brute des ENA après la réalisation de l'extraction des RS. L'objectif principal de ce post-traitement est de préparer les balises de RS reconnues pour qu'elles soient interrogées facilement par les

requêtes des utilisateurs lors de la recherche d'information pertinentes. Autrement dit, une requête posée pour chercher une ENA bien déterminée ne s'effectue pas sur sa forme annotée mais se compare avec sa forme brute. Pour réaliser ce post-traitement, nous avons créé deux graphes que nous présentons dans les figures suivantes.

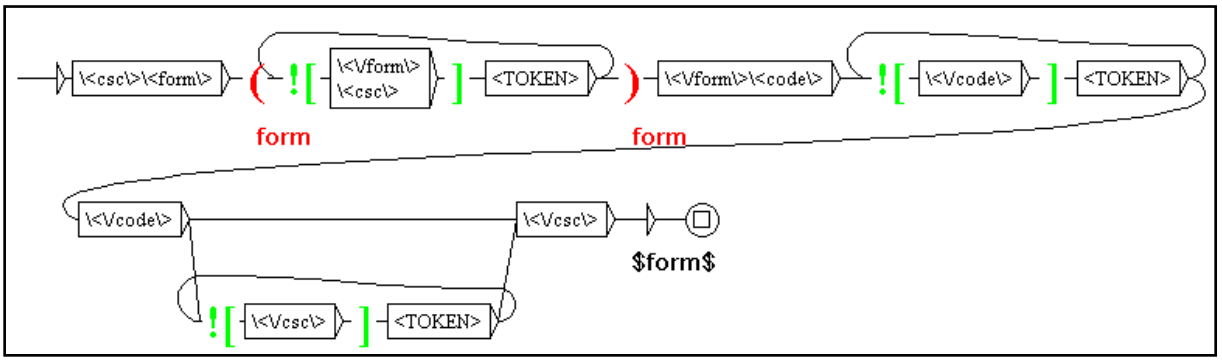


Figure 59. Graphe de récupération de la forme brute d'une ENA sans type

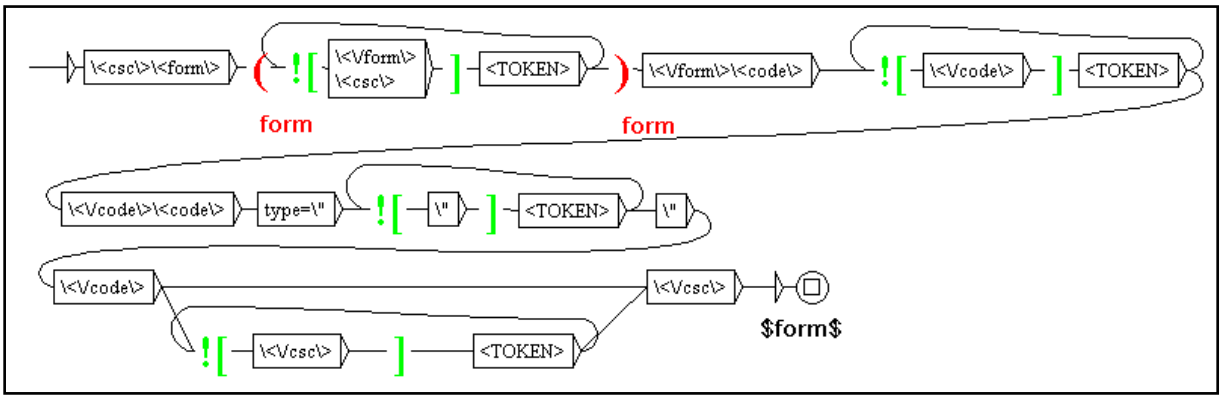


Figure 60. Graphe de récupération de la forme brute d'une ENA avec type

Les figures 59 et 60 décrivent les transducteurs de la récupération de la forme brute des ENA. Le principe de ces deux transducteurs rassemble à celui associé aux transducteurs dédiés à la normalisation. En fait, ces transducteurs exploitent également la version XML des textes générés après l'application du système ASRextractor. Rappelons que cette version en XML est générée grâce à l'outil CasSys. A partir de la forme balisée d'une ENA, les transducteurs lisent les deux balises <form><code> puis ils fixent une condition d'arrêt qui est leur fermeture selon </form> et </code>. La balise <form> stocke la valeur de l'ENA cherchée dans une variable appelée aussi form qui sera appelée dans le dernier nœud entre deux $.

Les transducteurs que nous avons établis doivent se regrouper dans une cascade fonctionnant selon deux modes précis pour réaliser l'objectif. Le premier mode doit remplacer les anciennes annotations d'ENA par leur forme brute. Pour cette raison, nous cochons le mode « replace ». De plus, ces transducteurs s'appliquent jusqu'au remplacement de toutes les occurrences des

ENA. Donc, nous cochons le mode appelé « Until Fix Point ». Dans la figure suivante, nous présentons la forme de cette cascade.

#	Disabled	Name	Merge	Replace	Until Fix Point	Generaliz...
1	☐	balisageTypeInverseNrm.fst2	☐	☑	☑	☐
2	☐	balisageInverseNrm.fst2	☐	☑	☑	☐

Figure 61. Cascade de récupération de la forme brute d'ENA

La figure 61 montre l'ordre de passage des deux graphes de traitement des balises contenant un élément type et des balises sans un élément type. Au sein de la nouvelle cascade réalisant le post-traitement déjà décrit, le mode « Until Fix Point » permet d'appliquer le graphe sélectionné jusqu'à ne plus avoir de changements dans le texte traité.

8. Conclusion

Dans le présent chapitre, les implémentations des systèmes proposés CasANER et ASRextractor ont été réalisées grâce à Unitex en tirant profit de ses fonctionnalités avancées. Etant donné que les systèmes proposés se basent sur des cascades de transducteurs, nous avons utilisé l'outil CasSys intégré sous Unitex pour les générer selon différents modes de passage. Notamment, nous avons pu paramétrer les systèmes proposés pour assurer leur performance au niveau analyse et annotation des textes traités. Dans ce sens, les textes exploités ont subi des prétraitements accomplis via Unitex. Vu que le rendement des transducteurs élaborés nécessite les entrées des dictionnaires, nous avons réussi à améliorer leurs couvertures et à les enrichir à travers une banque d'arbres syntaxiques. D'ailleurs, cette phase d'enrichissement s'appuie sur une étude complète. Le point important abordé par ce chapitre est la description du module de normalisation pour transformer l'annotation des ENA reconnues par le système CasANER à l'annotation conforme à la norme TEI. Pour exploiter la sortie des systèmes proposés, nous avons proposé un post-traitement associé aux ENA figurant dans les RS. Ce post-traitement vise à récupérer la forme brute des ENA dont l'objectif est de faciliter l'exploitation des RS pour enrichir les dictionnaires d'ENA.

Par la suite, les systèmes proposés doivent être expérimentés sur un corpus de test extrait à partir de la Wikipédia arabe. Cette expérimentation donne l'opportunité de visualiser les résultats et les analyser pour faciliter la phase d'évaluation. D'ailleurs, cette évaluation va se baser sur les métriques de performance afin d'effectuer une interprétation profonde.

Chapitre 8 : Expérimentation et évaluation des systèmes CasANER et ASRextractor

L'évaluation des systèmes CasANER et ASRextractor va favoriser leur amélioration après une phase d'expérimentation sur un corpus de test extrait de la Wikipédia arabe. Pour ce faire, nous choisissons les mesures de performance qui sont les plus utilisées pour cerner les lacunes via des calculs à la base des résultats obtenus. Les formules de ces mesures sont modifiables selon le domaine étudié par un système proposé comme la REN dont le calcul s'articule autour des EN. L'évaluation du système CasANER peut s'étendre pour qu'elle se base sur une comparaison avec un autre système de REN existant. D'ailleurs, cette évaluation n'est pas une tâche triviale vu qu'elle nécessite des traitements à faire surtout la mise en correspondance des annotations des ENA fournies par les deux systèmes. Diverses difficultés rencontrées peuvent se présenter aussi dans le cadre de la comparaison. Autrement dit, le système CasANER doit être passé sur le même corpus utilisé par le système candidat sachant que ce corpus admet qu'une version annotée. Dans ce cas, nous envisageons une phase de filtrage de ce corpus pour récupérer son format brut. De plus, nous devons adopter les deux annotations afin de faciliter la comparaison.

Dans le chapitre courant, nous présentons la phase d'expérimentation associée aux systèmes CasANER et ASRextractor. Dans la même phase, nous proposons la visualisation des résultats de normalisation de l'annotation des ENA en TEI et les résultats des post-traitements effectués sur les RS. Après, nous passons à l'évaluation des deux systèmes CasANER et ASRextractor pour prouver leurs performances. En premier lieu, nous les évaluions globalement à travers les métriques Rappel, Précision et F-mesure pour les résultats obtenus. Ensuite, nous raffinons ce genre d'évaluation pour qu'il se décompose de l'évaluation de CasANER par catégorie d'ENA et celle de ASRextractor par type de RS. En second lieu, nous comparons CasANER avec un système de REN basé sur l'approche statistique. En fait, cette partie nécessite un traitement spécifique sur le corpus exploité pour réaliser la comparaison. En se basant sur cette comparaison, nous discutons les problèmes rencontrés ainsi que les avantages du processus de REN fait par notre système CasANER.

1. Phase d'expérimentation

La phase d'expérimentation des systèmes implémentés est une initiation à leur évaluation. Cette phase consiste à appliquer ces systèmes sur le corpus de test qui a été collecté parallèlement à la collection du corpus d'étude. Nous rappelons que le nouveau corpus a été obtenu aussi grâce à l'outil Kiwix. De plus, nous devons mentionner que le corpus de test n'a été exploité dans aucune phase des travaux élaborés. En fait, la visualisation des résultats obtenus est assurée par l'éditeur du texte Notepad++ v6.8.8 qui a facilité la lecture des textes arabes et l'analyse des

balises qui entourent les ENA et les RS. Dans la section suivante, nous décrivons des extraits issus de l'expérimentation associée à chaque travail implémenté.

1.1. Expérimentation du système CasANER

L'application du système CasANER s'effectue sur le corpus de test brut qui a subi à son tour un prétraitement consistant à le segmenter et à éliminer les liens internes indésirables. Après ce prétraitement, le système CasANER a généré une nouvelle version de ce corpus dont les ENA sont reconnues et annotées à travers les accolades. Pour montrer le résultat obtenu, nous proposons un extrait d'un texte arabe appartenant au corpus de test.

Figure 62. Extrait d'un fichier de sortie de CasANER

La figure 62 décrit deux extraits différents d'un même texte arabe associé au nom d'une personne célèbre appelée « دوللي شاهين/ Dolly Chahine ». Le premier extrait représente la version brute qui était initialement l'entrée du système CasANER avant le prétraitement. Cependant, le deuxième est une sortie générée par notre système après la segmentation et la suppression des liens internes qui vont empêcher le processus. Pour cette raison, nous remarquons la présence des balises {S} et [Link] ajoutés par les graphes élaborés précédemment. D'après le deuxième extrait, nous trouvons que le système CasANER reconnaît des ENA et les annote selon les trois catégories « persName » et « placeName » et « Date ». De plus, cet extrait montre l'annotation qui est détaillée et raffinée.

1.2. Expérimentation du système ASRextractor

L'expérimentation du système ASRextractor se base sur la sortie du système CasANER vu qu'il génère un corpus contenant des ENA reconnues et annotées. De plus, le système ASRextractor profite de l'annotation sous la forme des accolades puisque ce type d'annotation rend l'ENA reconnue exploitable comme un trait d'un dictionnaire. Pour observer les RS extraites et annotées en TEI, nous proposons l'extrait d'un texte faisant partie du corpus de test annoté.

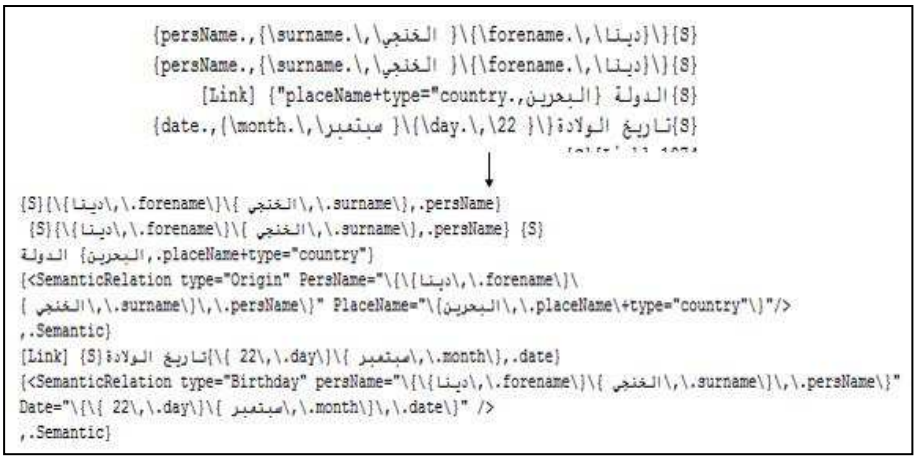


Figure 63. Extrait d'un fichier de sortie de ASRextractor

La figure 63 montre que le système ASRextractor a pu extraire deux types de RS à partir des 4 premières lignes du texte. Les deux types extraits sont « Origine » et « Date de naissance » qui concerne un nom de personne. D'après cette figure, nous constatons que les composants des RS sont bien organisés et représentés à travers les éléments exploités comme type, persName, Date et OriginPlace.

1.3. Expérimentation de la cascade de normalisation

L'application de la cascade de normalisation fonctionne bien lors de son application sur le corpus de test, plus précisément en profitant de la sortie de CasANER. Dans la figure suivante, nous proposons une illustration de la sortie de cette cascade de normalisation.

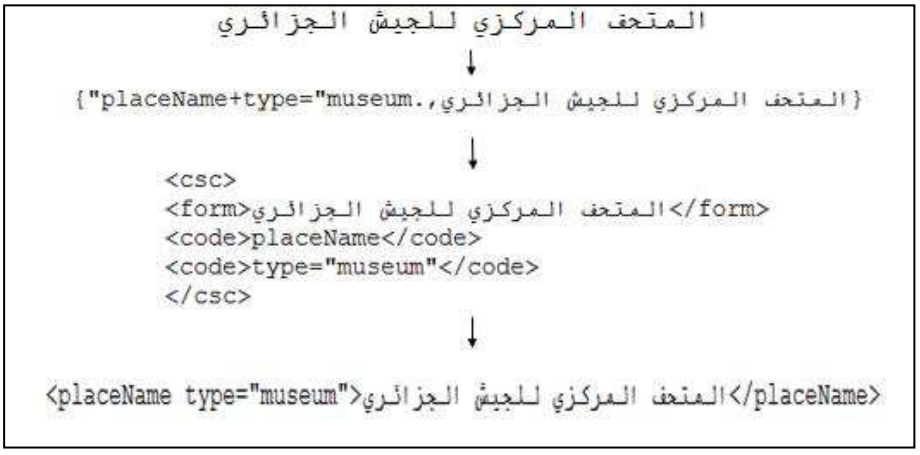


Figure 64. Annotation normalisée d'une ENA

La figure 64 illustre la transformation de l'annotation d'une ENA ayant la catégorie nom de lieu relatif plus précisément un nom de musée. Dans la première partie, l'ENA entre accolades a été reconnue et annoté via CasANER. Puis, nous avons sa traduction selon le fichier XML généré par CasSys qui était l'entrée de la cascade de normalisation pour produire l'ENA structurée dans la troisième partie. En fait, la forme d'annotation finale est associée à la TEI.

1.4. Expérimentation de la cascade de récupération de la forme brute des ENA

Après l'application du système ASRextractor, nous avons pensé à préparer le corpus de test pour son exploitation par d'autres applications de TAL. Cette fois, nous n'allons pas exploiter le corpus de test avec l'annotation des accolades mais nous allons profiter de sa version en XML. Cette dernière se génère automatiquement grâce à l'outil CasSys. De plus, les fichiers en XML se trouvent dans le même répertoire des fichiers texte issus du système ASRextractor. Après l'application de la cascade de récupération de la forme brute des ENA dans les RS extraites, nous proposons deux extraits d'un même texte du corpus de test pour monter le résultat obtenu.

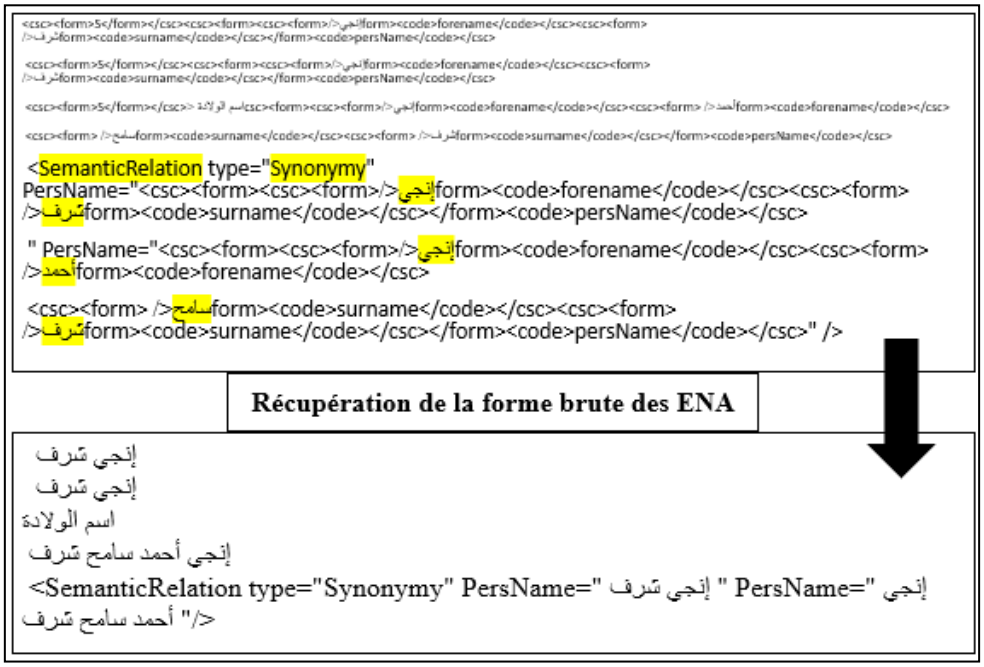


Figure 65. Résultat de récupération de la forme brute des ENA

La figure 65 décrit un extrait d'un texte appelé « إنجي شرف / Enji Charaf » avant et après l'application de la cascade de récupération de la forme brute d'ENA. Nous constatons que toutes les ENA ont été traitées par cette cascade. En fait, la nouvelle forme des RS offre une représentation claire et visible. Les résultats obtenus favorisent dans ce cas l'utilisation de notre corpus de test pour enrichir des dictionnaires d'ENA ou d'autres corpus arabes par les RS.

L'application de la cacade de normalisation et celle de récupération de la forme brute des ENA n'ont pas une influence sur la qualité des systèmes élaborés. En fait, elles permettent juste de traiter la sortie de ces systèmes pour les exploiter dans d'autres applications de TAL comme l'enrichissement des dictionnaires d'ENA.

2. Evaluation de CasANER pour toutes les catégories

L'évaluation de notre système CasANER pour la reconnaissance et l'annotation des EN est un processus permettant de prouver sa fiabilité. Pour cette raison, nous exploitons initialement les mesures de performance (précision, rappel et F-mesure) pour évaluer toutes les catégories ensembles. Par la suite, nous effectuons l'évaluation par catégorie pour cerner les failles de ce système. Avant de présenter le tableau récapitulatif, nous donnons les formules exploitées pour effectuer le calcul.

$$Rappel = \frac{Nombre\ d'ENA\ reconnues\ justes}{Nombre\ d'ENA\ reconnues\ justes\ et\ d'ENA\ non\ reconnues}$$

$$Précision = \frac{Nombre\ d'ENA\ reconnues\ justes}{Nombre\ d'ENA\ reconnues}$$

$$F - mesure = \frac{2 * Rappel * Précision}{Rappel + Précision}$$

En appliquant ces formules sur les résultats obtenus après le test, nous obtenons les valeurs présentées dans le tableau ci-dessous.

Tableau 15. Evaluation de CasANER avec les métriques de performance

Corpus	Rappel	Précision	F-mesure
95 378 tokens	0.91	0.92	0.91

Le tableau 15 montre que CasANER touche une précision de 92%, un rappel de 91% et un F-mesure de 91% pour la reconnaissance des ENA. Par conséquent, nous constatons que les résultats obtenus sont très motivants. En fait, le nombre d'ENA reconnue en erreur cause la valeur de rappel obtenue. Les erreurs générées lors de processus de REN réalisée par CasANER sont dues au fait que la couverture du dictionnaire est insuffisante dans certains cas. Celle-ci doit être améliorée. En outre, la performance de la REN augmente si la couverture des dictionnaires est enrichie. Toutefois, il existe des types d'erreurs pouvant être causées par la structure des articles de la Wikipédia arabe. Par exemple, il y a des ENA ayant des mots de déclenchement incomplets, tels que l'ENA « انتفاضة الصدر » au lieu de « انتفاضة الصدر / la révolution de Al-Sader » et le mot déclencheur « محافة » remplaçant le mot « محافظة/ ville » précédant les noms de lieux absolus plus précisément les noms de villes. Il faut mentionner aussi que les prépositions jouent un rôle très important dans la REN vu qu'elles précèdent la majorité des ENA. Néanmoins, dans la Wikipédia arabe il existe plusieurs erreurs de frappe au niveau des prépositions, telles que « قي » au lieu de « في / dans ».

3. Evaluation de CasANER par catégorie

Pour tester la performance de CasANER par catégories, nous évaluons chacune de ces dernières en utilisant également des mesures déjà définies. Cette décomposition nous aide à déterminer les catégories nécessitant une amélioration.

Tableau 16. Evaluation de CasANER par catégorie

	Date	**Nom de personne**	**Evènement**	**Nom de lieu**	**Organisation**
Rappel	0.78	0.87	0.92	0.95	0.99
Précision	0.81	0.95	0.96	0.94	0.97
F-mesure	0.79	0.90	0.93	0.94	0.97

Le tableau 16 prouve que notre système CasANER excelle, en particulier, dans la reconnaissance des catégories Événement, Nom de lieu et Organisation. Le taux de reconnaissance de la catégorie Nom de personne est intéressant puisque nous avons utilisé trois types de transducteurs d'analyse, de filtrage et de généralisation d'étiquetage, ce qui nous aide à reconnaître diverses formes d'ENA assignées à cette catégorie. Les transducteurs de filtrage ont permis de récupérer les ENA qui n'étaient séparées que par des séparateurs imprévus comme celui de segmentation. L'utilisation des transducteurs de généralisation d'étiquetage permet d'éviter des problèmes d'ambiguïté. Par exemple, avant leur utilisation nous étions face à une étape de prédiction si un mot est un prénom ou un nom de famille et non un adjectif ou un nom commun. Ce genre de problème se pose plus que les autres problèmes liés à la catégorie Nom de personne. Avoir de bonnes valeurs pour trois catégories principales (*Evénement, Nom de lieu* et *Organisation*) sont justifiées au fait qu'elles possèdent un nombre très important de règles d'extraction identifiées par rapport au reste de catégories. D'ailleurs, le nom de lieu possède un nombre remarquable de sous catégories. Cependant, la reconnaissance d'une année sans mots déclencheurs et des prépositions provoque les valeurs obtenues pour la catégorie *Date*. Cette forme de Date a participé à baisser la précision de reconnaissance de cette catégorie. En outre, les erreurs peuvent être trouvées dans les prépositions, qui peuvent jouer le rôle d'indicateur pour déterminer les limites ENA, telles que « قي » au lieu de « في/ dans ».

4. Evaluation de ASRextractor pour tous les types de RS

Après la présentation de la cascade de transducteurs composant notre système ASRextractor, nous passons à évaluer la qualité d'extraction des RS après son passage sur le corpus de test. L'évaluation d'ASRextractor va nous permettre de vérifier la suffisance des règles d'extraction considérées pour détecter les RS entre les ENA. Cette évaluation va nous donner également

l'opportunité de maîtriser les points forts de ce système et d'avoir une chance de remédier les points faibles.

Rappelons que l'évaluation a été faite aussi manuellement. La qualité d'extraction s'est effectuée à base des mesures de précision, de rappel et de la F-mesure telles qu'elles sont définies comme suit :

$$Rappel = \frac{Nombre\ de\ RS\ reconnues\ justes}{Nombre\ de\ RS\ reconnues\ justes\ et\ de\ RS\ non\ reconnues}$$

$$Précision = \frac{Nombre\ de\ RS\ reconnues\ justes}{Nombre\ de\ RS\ reconnues}$$

$$F - mesure = \frac{2 * Rappel * Précision}{Rappel + Précision}$$

En appliquant ces formules sur les résultats obtenus après le test, nous obtenons les mesures présentées dans le tableau suivant.

Tableau 17. Evaluation de ASRextractor avec les métriques de performance

Corpus	Rappel	Précision	F-mesure
246 001 tokens	0.83	0.9	0.86

En visualisant les résultats présentés dans le tableau 17, nous remarquons leur satisfaction à partir de la performance des RS identifiées. L'évaluation manuelle sur le corpus de test montre qu'ASRextractor est capable de couvrir la majorité des RS reliant les ENA avec une précision de 0,9, un rappel de 0,83 et une F-mesure d'environ 0,86. Nous voulons signaler quelques erreurs qui ont contribué à cette valeur de Rappel.

La raison principale est due aux erreurs générées par le système CasANER vu qu'il existe certaines ENA non reconnues et annotés. D'autres anomalies sont dues au manque des contraintes suffisantes dans les règles d'extraction. Autrement dit, il y de nouveaux chemins d'extraction de RS qui ne sont pas étudiés par notre système ASRextractor. Malgré que les chemins existants aient souffert de la couverture du dictionnaire sémantique créé pour guider le processus d'extraction. D'où vient la nécessité d'augmenter cette couverture afin de favoriser l'extraction des RS non reconnues.

Il ne faut pas oublier que l'amélioration doit toucher aussi les dictionnaires d'Unitex vu que les chemins d'extraction sont sous forme de segments syntaxiques qui en ont besoin. Par exemple, le mot « العاصمة / capitale » était un indicateur d'une RS d'accessibilité inexistant dans le dictionnaire des noms communs. Par conséquent, cette RS n'a pas pu être détectée. Finalement, nous avons toujours le même souci lié à la structure de la Wikipédia plus précisément aux prépositions.

Pour bien expliquer les problèmes liés à la structure de la Wikipédia, nous proposons l'exemple illustré dans la figure 66. Dans cette figure, la RS de méronymie est non extraite à cause d'une erreur d'écriture d'un indicateur sémantique. En fait, la méronymie ne se détecte pas car la préposition qui est l'élément sémantique central est écrite en erreur.

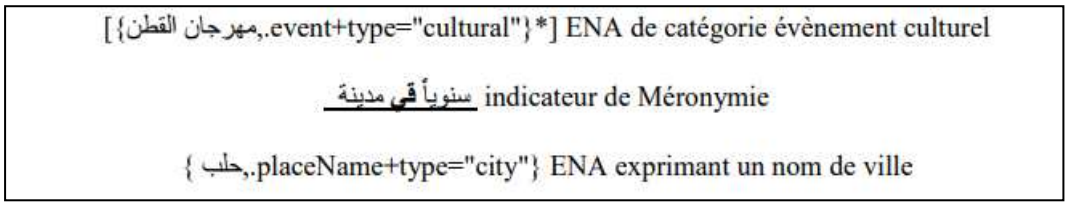


Figure 66. Exemple d'erreur liée à la structure de la Wikipédia

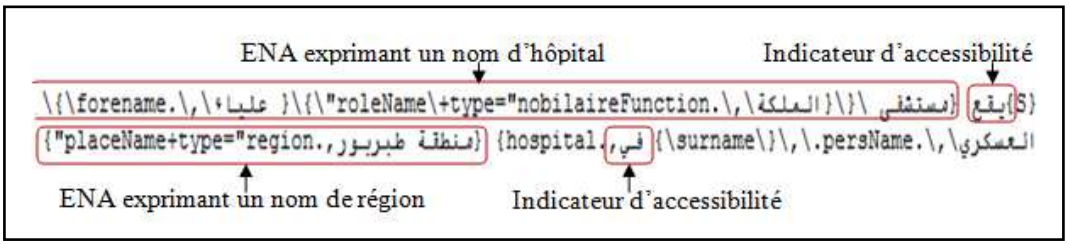


Figure 67. Exemple d'erreur générée par le système CasANER

Revenons aux problèmes des ENA non reconnues par le système CasANER. Dans l'exemple illustré dans la figure 67, une RS d'accessibilité se déclenche à travers un indicateur sémantique ayant la forme d'un verbe « يقع / se situe ». Le système CasANER reconnaît le nom de l'hôpital en erreur et inclut la préposition « في / dans » au sein de l'ENA reconnue. Or, cette préposition est censée indiquer une RS d'accessibilité entre ce nom d'hôpital et un nom de région.

5. Evaluation d'ASRextractor par type de RS

Le système ASRextractor a besoin d'une seconde évaluation qui va toucher chaque type de RS appartenant à la typologie établie dans le chapitre quatre. Cette nouvelle sorte d'évaluation permet de découvrir les types de RS ayant des règles d'extraction insuffisantes par rapport à leur présence dans les corpus étudiés. En outre, les types de RS ayant une valeur de rappel faible sont les plus fréquentes dans les textes traités. Pour ce faire, nous appliquons aussi les mesures de performance afin de proposer ultérieurement des améliorations. Le tableau suivant décrit les résultats d'évaluation que nous avons effectués par type de RS. En fait, nous organisons les valeurs obtenues en nous basant sur celles liées à la précision à partir de la valeur moyenne jusqu'à la meilleure.

Tableau 18. Evaluation d'ASRextractor par type de RS

	Précision	**Rappel**	**F-mesure**
Age	0.6	1	0.75
Date politique	0.68	0.84	0.75
Année de fondation	0.8	0.8	0.8
Fonctionnel	0.72	0.66	0.68
Appartenance	0.87	0.87	0.78
Accessibilité	0.89	0.69	0.77
Synonymie	0.9	1	0.94
Proximité	0.92	1	0.95
Origine	0.94	0.89	0.91
Méronymie	0.95	1	0.97
Famille	1	0.66	0.79
Profession	1	0.82	0.9
Nationalité	1	0.85	0.91
Date de naissance	1	0.92	0.95

Le tableau 18 prouve qu'ASRextractor excelle en quatre types de RS qui sont Date de décès, Equivalence date, Nationalité et Origine ayant un F-mesure égal à 1. Ces types déjà mentionnés sont extraits à partir d'un ensemble d'expressions régulières exhaustives. De plus, ces cinq types de RS apparaissent dans des textes ayant une structure en commun. Notamment, ils sont toujours exprimés par des segments courts ce qui rend leur extraction plus facile. Concernant le reste des types, les valeurs obtenues nous poussent à réviser les expressions régulières créées pour augmenter les champs de détection pour chaque type. En revanche, ces types peuvent apparaître dans des nouveaux contextes non analysés lors de notre étude linguistique.

6. Évaluation CasANER sur ANERcorp et comparaison avec ANERsys

La seconde évaluation de CasANER encourage à tester sa performance sur un nouveau corpus. De plus, elle permet d'analyser les résultats obtenus et de cerner les catégories d'ENA nécessitant une amélioration. Cependant, le corpus exploité ne doit pas être utilisé lors de l'élaboration du système CasANER. Pour cette raison, nous comptons utiliser l'ANERcop qui est le corpus le plus exploité dans la REN associée à la langue arabe. Après le passage de CasANER sur le ANERcorp, nous allons le comparer à un système de REN qui est basé sur l'approche statistique, appelé ANERsys avec le même corpus. Avant d'entamer ces deux phases d'évaluation et comparaison, nous passons à la présentation du corpus ANERcorp qui est préalablement annoté.

6.1. Présentation du corpus ANERcorp

ANERcorp est un corpus accessible librement et construit par [Benajiba et al., 2007]. ANERcorp a été collecté à partir de plusieurs sources textuelles afin d'obtenir un corpus aussi

généralisé que possible. Ce corpus a été collecté comme un corpus d'étude et de test pour le système appelé ANERsys élaboré par les auteurs déjà mentionnés. ANERsys est un système de REN basé sur l'approche statistique. Rappelons qu'ANERcorp contient 316 articles de presses et il possède au total plus que 150 000 mots annotés lors du processus de REN. Chaque mot a été annoté suivant un balisage défini dans le tableau 19 ci-dessous.

Tableau 19. Balises d'annotation utilisées par ANERsys

Balise	Signification
B-PERS	Début d'un nom de personne
I-PERS	Intérieur d'un nom de personne
B-LOC	Début d'un nom de lieu
I-LOC	Intérieur d'un nom de lieu
B-ORG	Début d'un nom d'organisation
I-ORG	Intérieur d'un nom d'organisation
B-MISC	Début d'une EN n'appartenant pas aux trois catégories précédentes
I-MSC	Intérieur d'une EN n'appartenant pas aux trois catégories précédentes
O	Annotation d'un mot quelconque

ANERcorp est associé également à trois types de dictionnaires appelés selon les auteurs « ANERgazet ». Ces derniers sont construits manuellement en utilisant des ressources textuelles provenant du Web. Le premier dictionnaire stocke les noms de lieu et possède 1 950 entrées collectées à partir de la Wikipédia arabe. La couverture du dictionnaire des noms de personne est plus large avec un nombre d'entrées égal à 2 100. Leur collection a été faite aussi via Wikipédia, plus d'autres sites web non déterminés. Le dernier dictionnaire, rassemblant les noms d'organisation, contient un nombre réduit d'entrées, égal à 262.

Avant d'entamer l'explication du processus d'évaluation que nous proposons, nous devons mentionner que les « ANERgazet » ne sont utilisées dans aucune phase d'élaboration, ni de CasANER, ni d'ASRextractor. En fait, nos dictionnaires sont plus larges en termes de couverture. Dans la sous-section suivante, nous expliquons le principe d'évaluation de CasANER sur l'ANERcorp et sa comparaison avec un autre système.

6.2. Principe de l'évaluation et la comparaison

L'évaluation de notre système CasANER consiste à le passer sur le corpus ANERcorp. Cette évaluation n'est pas une tâche facile puisqu'elle se base sur plusieurs étapes. De plus, cette évaluation est un préliminaire à la réalisation de la comparaison de deux systèmes.

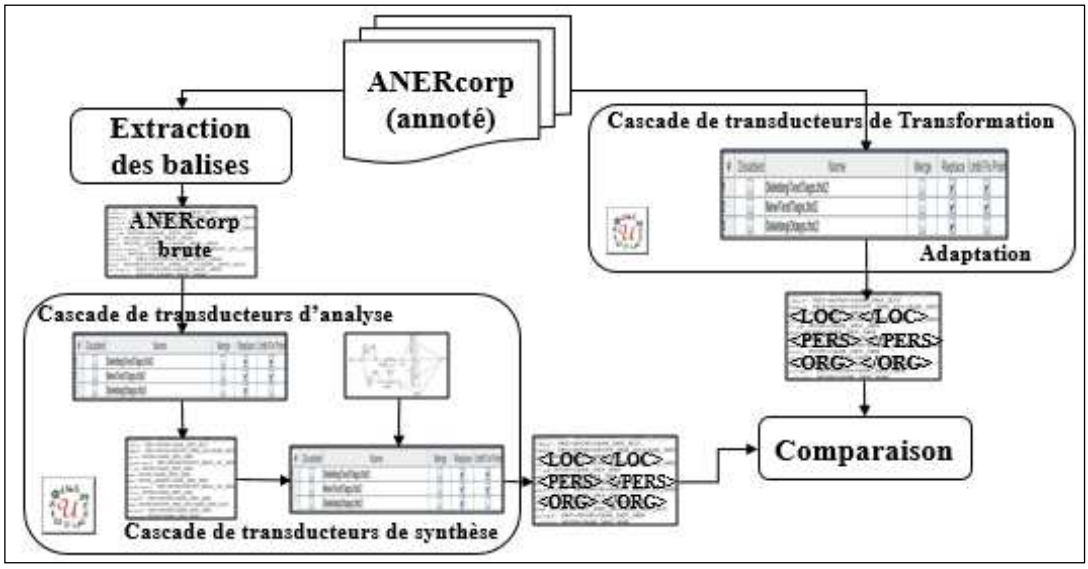


Figure 68. Principe de l'évaluation du système CasANER sur le ANERcorp

La figure 68 montre une architecture qui résume le principe de l'évaluation de la performance du système CasANER sur le corpus ANERcorp et sa comparaison avec un système de REN statistique. D'après la figure, nous constatons que ces deux phases déjà mentionnées nécessitent un traitement puisque le corpus ANERcorp est annoté. Le traitement à effectuer se compose de deux parties principales. La première partie récupère la forme brute d'ANERcorp, puis celle de CasANER. La deuxième partie effectue un processus d'adaptation pour rendre l'annotation au sein de corpus ANERcorp conforme à celles de notre système CasANER. Dans ce qui suit, nous expliquons chaque étape illustrée.

6.2.1. Prétraitements effectués sur ANERcorp

Pour faciliter le passage de CasANER sur l'ANERcorp nous allons effectuer certains prétraitements. Ce processus de prétraitement se compose de deux étapes principales dont la première vise à récupérer la version brute d'ANERcorp. Tandis que la deuxième étape est dédiée à l'adaptation des balises d'annotation des deux systèmes à comparer, CasANER et ANERsys. Pour expliquer le principe, nous proposons le schéma ci-dessous.

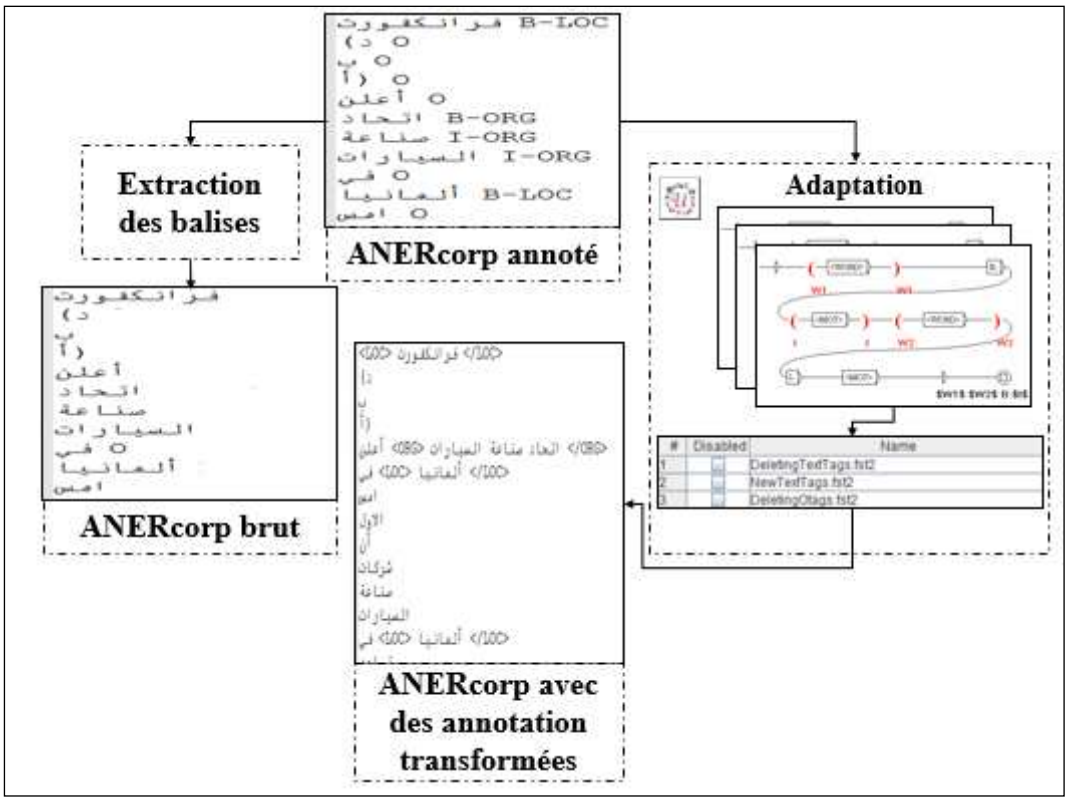


Figure 69. Prétraitements effectués sur ANERcorp

La figure 69 explique les étapes de prétraitement que nous effectuons sur l'ANERcorp. La première étape consiste à extraire les balises afin de les éliminer et d'avoir une version brute d'ANERcorp. Comme il est illustré dans cette figure, le corpus ne contient aucune balise. Cette extraction est simple et se réalise sur l'éditeur de texte Notepad++ avec un Ctrl+F. Après il faut chercher toutes les cooccurrences de chaque balise (voir tableau 7) et les remplacer par le vide. La deuxième étape qui vise à adapter les annotations d'ANERcorp à celle de CasANER n'est pas une tâche facile. En outre, elle est composée par d'autres sous-étapes. A partir de la version annotée d'ANERcorp nous avons créé trois graphes d'adaptation qui seront regroupés au sein d'une cascade de transducteurs fournissant la sortie souhaitée. Afin de clarifier les sous-étapes d'adaptation, nous présentons dans ce qui suit les trois graphes déjà cités et la forme de la cascade ainsi leur mode de passage. Nous commençons par la description de nos graphes.

6.2.2. Création des graphes d'adaptation d'annotation

L'analyse de la version annotée d'ANERcorp est une phase très intéressante. Cela nous permet de comprendre l'enchaînement des balises. Nous avons essayé d'exploiter notre analyse pour atteindre l'objectif qui consiste à remplacer les balises existantes dans ce corpus par de nouvelles balises ayant la forme ENA. Par exemple, nous remplaçons le B-PERS et I-PERS par <PERS> et </PERS>. Au départ, cette transformation est compliquée vu qu'une ENA n'existe pas dans la même ligne mais elle est répartie sur plus que deux lignes lorsqu'elle est composée.

Nous avons réussi à créer des graphes qui nous ont permis d'avoir un résultat adéquat sans toucher le contenu des ENA.

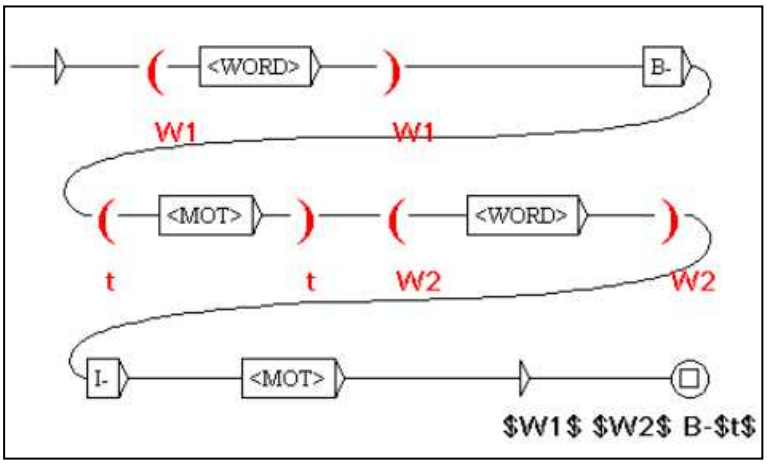


Figure 70. Graphe de transformation de l'annotation d'une ENA composée

La figure 70 illustre le chemin de transformation de l'annotation associée à une ENA annotée dans l'ANERcorp. Autrement dit, cette ENA est annotée avec B-catégorie et I-catégorie que nous avons expliqué auparavant. Prenons un exemple dans le corpus, « براند جوتشالك » est une ENA décrivant un nom de personne était divisé sur deux lignes. La première ligne contient « براند B-PERS » mais la deuxième contient « جوتشالك I-PERS ». Après le passage de ce graphe illustré l'annotation devient comme suit : براند جوتشالك B-PERS. Ce graphe est capable de reconnaître une ENA distribuée sur plus que 2 lignes.

La transformation d'annotation pour une ENA non composée n'est pas une tâche triviale. Autrement dit, il faut traiter celle-là tout en évitant de récupérer un mot quelconque. Ce principe est bien illustré dans le graphe suivant.

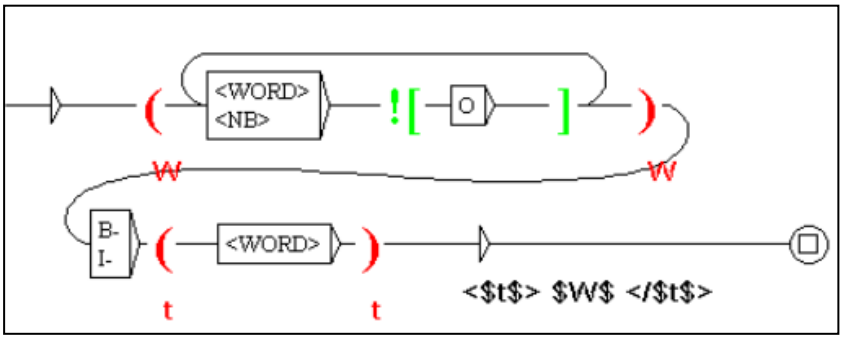


Figure 71. Graphe d'une deuxième transformation d'annotation d'une ENA

La figure 71 décrit la transformation de l'annotation d'une ENA traitant plusieurs cas. Le premier se présente lors une ENA qui contient qu'une seule composante. En outre, cette ENA peut exprimer un nom de lieu relatif ou un nom d'organisation réduit ou encore un seul élément d'un nom de personne. Un autre cas traité par ce graphe est le résultat du graphe précédent

c'est-à-dire lorsque l'ENA a été regroupée dans la même ligne et qu'il faut avoir une nouvelle annotation. Dans ce graphe, nous avons utilisé la notion de contexte négatif qui a été expliquée dans d'autres sections. Ce contexte négatif essaie de se rassurer que l'annotation traitée est bel et bien représentée par une balise caractérisant une ENA et non un mot quelconque. Par exemple, après le passage de ce graphe illustré l'annotation de l'ENA براند جوتشالك B-PERS devient comme suit : <PERS> براند جوتشالك </PERS>. Prenons un autre exemple qui fait partie aussi de corpus, le nom de lieu ألمانيا était équipé par la balise B-LOC et grâce à ce graphe l'annotation devient <LOC> ألمانيا </LOC>.

Etant donné que tous les mots dans ANERcorp sont annotés, nous avons pensé à créer un graphe pour éliminer la balise O qui annote tout mot autre qu'une ENA. Nous illustrons sa forme dans la figure ci-dessous.

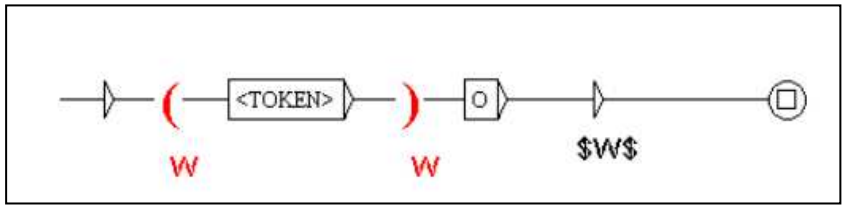


Figure 72. Graphe d'élimination de l'annotation d'un mot quelconque

La figure 72 illustre la forme d'un graphe dédié à éliminer la balise O annotant des mots quelconques. Le mot dans ce cas est considéré comme un token suivi par la balise déjà mentionnée. Par exemple, le mot « لعام O » après le passage de ce graphe, il devient « لعام ».

Tous ces graphes contiennent la notion de variable car ils permettent de remplacer l'ancienne annotation par une nouvelle. De plus, les ENA ont été cherchées en tant que mot représenté par le trait <WORD>. Par contre, le troisième graphe utilise le trait <TOKEN> car le mot cherché peut être un signe de ponctuation. De plus, nous n'avons pas fait de graphes de filtrage car nous avons constaté que nos graphes répondent bien à réaliser notre objectif.

6.2.3. Création de la cascade d'adaptation de l'annotation

Les trois graphes déjà expliqués ont formé une nouvelle cascade prenant en entrée ANERcorp annoté et génèrent en sortie une nouvelle forme d'annotation de ce corpus. Dans la figure suivante, nous montrons la forme de cette cassade d'adaptation.

#	Disabled	Name	Merge	Replace	Until Fix Poin
1	☐	DeletingTextTags.fst2	☐	☑	☑
2	☐	NewTextTags.fst2	☐	☑	☑
3	☐	DeletingOtags.fst2	☐	☑	☐

Figure 73. Cascade d'adaptation de l'annotation d'ANERcorp

La figure 73 montre que l'appel des graphes au sein de la cascade d'adaptation s'effectue dans un ordre précis. Cet ordre a été expliqué dans la présentation de ces trois graphes dans les figures précédentes pour montrer qu'il n'est pas aléatoire. Les modes passage de cette cascade est « replace » pour le remplacement d'annotation et « Until fix point » pour traiter toutes les occurrences jusqu'à avoir un texte inchangé.

Adapter l'annotation d'ANERcorp à celle fournie par CasANER est une étape importante. Néanmoins, cette dernière donne naissance à une nouvelle étape qui consiste à adopter l'annotation de CasANER avec la nouvelle forme générée par la cascade d'adaptation. Cette nouvelle étape sera réalisée après avoir passé notre système sur ANERcorp brute. Ce passage permettra d'annoter toutes les catégories figurant dedans. Par contre, l'évaluation et la comparaison que nous voulons réaliser ne se fera qu'avec les trois catégories *Nom de personne, Nom de lieu* et *Nom d'organisation*.

6.3. Evaluation et comparaison

Pour évaluer la performance de CasANER, nous le passons sur un corpus qui n'a pas été exploité dans aucune phase de ce système, ni dans les dictionnaires, ni dans les corpus d'étude et de test. Le passage de CasANER était suivi par la cascade de synthèse pour avoir une annotation normalisée. Nous illustrons les étapes de REN dans l'architecture suivante.

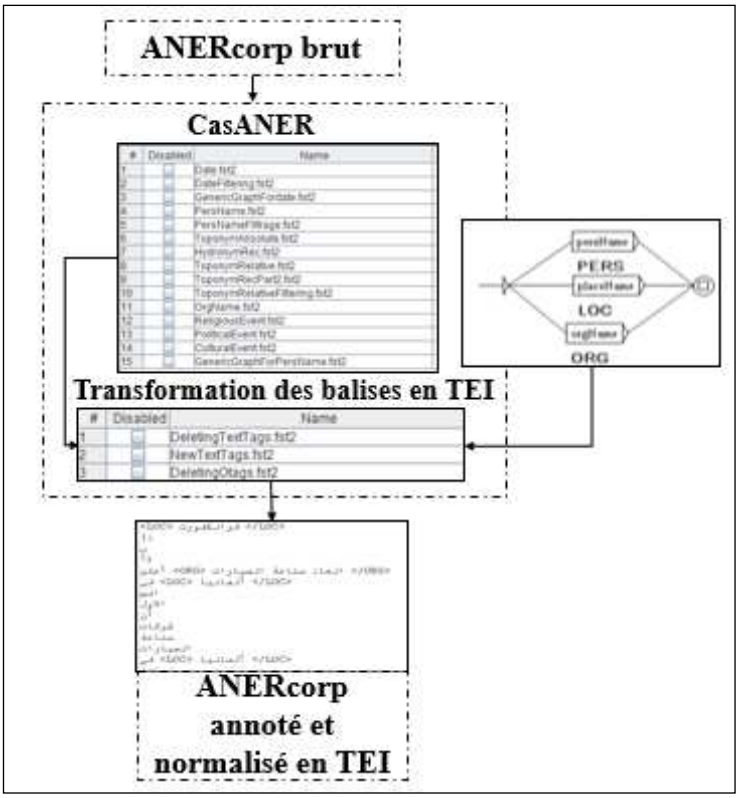


Figure 74. Architecture de la REN effectuée par CasANER sur ANERcop brut

La figure 74 décrit le processus de REN en se basant sur l'ANERcorp. Dans ce processus, CasANER fonctionne d'une façon ordinaire. En revanche, la différence réside lors de l'application de la cascade de synthèse pour normaliser l'annotation des ENA. En fait, nous avons rajouté un petit graphe pour avoir une sortie conforme avec celle de ANERsys. Cette modification va faciliter la comparaison après. Pour montrer la crédibilité des résultats de CasANER, nous rappelons que nous utilisons juste le corpus brut pour la REN et la version transformée en termes d'annotation pour la comparaison.

Le graphe qui a été rajouté à la cascade de synthèse ne touche que les trois catégories traitées par le système ANERsys. Les autres catégories auront le même principe d'annotation selon la norme TEI. C'est important de mentionner que la forme de trois balises pour respectivement *Nom de personne, Nom de lieu* et *Nom d'organisation* changent mais le niveau interne de raffinement sera toujours le même. En outre, chaque balise aura des imbrications et des éléments indiquant les sous-catégories traitées.

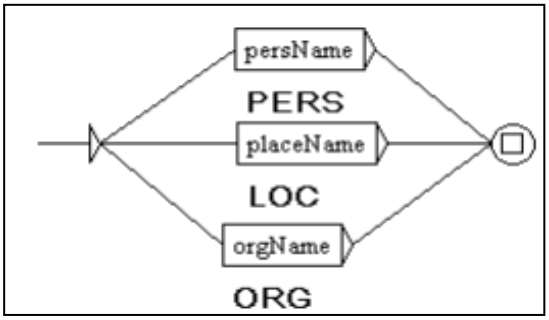


Figure 75. Graphe de transformation ajouté à la cascade de synthèse

La figure 75 montre le contenu du graphe de transformation. Ce graphe effectue une simple modification en gardant toujours la notion de balise (ouvrante et fermante). Par exemple, les balises <placeName> et </placeName> cernant une ENA de catégorie Nom de lieu seront remplacées par <LOC> et </LOC>. Comme nous l'avons déjà dit ce graphe sera intégré au sein de la cascade de synthèse à la troisième position. Il va être passé en mode replace comme elle indique la figure 76.

#	Disabled	Name	Merge	Replace	Until Fix Point
1	☐	balisageType.fst2	☐	☑	☑
2	☐	balisage.fst2	☐	☑	☑
3	☐	TransformTags.fst2	☐	☑	☐

Figure 76. Nouvelle forme de la cascade de normalisation

Après le passage de CasANER et la cassade de synthèse sur l'ANERcorp brut, nous évaluons les résultats obtenus à travers les mesures de performance qui sont le rappel, la précision et la F-mesure. En fait, nous adoptons la même définition de chaque mesure illustrée dans la sous-

section précédente dédiée à la REN. Pour cette raison, nous illustrons un tableau comparatif entre les résultats de CasANER et ceux fournis par ANERsys.

Tableau 20. Comparaison entre ANERsys et CasANER

	ANERSys			**CasANER en utilisant ANERcorp**		
	LOC	**PERS**	**ORG**	**LOC**	**PERS**	**ORG**
Rappel	0.78	0.41	0.31	0.71	0.81	0.58
Précision	0.82	0.54	0.45	0.66	0.76	0.55
F-mesure	0.80	0.46	0.36	0.67	0.78	0.63

Le tableau 20 décrit les valeurs obtenues par à notre système CasANER et ceux propres à ANERsys. D'après ce tableau, nous pouvons constater que CasANER excelle par rapport à l'autre système. En fait, notre système était plus efficace qu'ANERsys dans la reconnaissance de noms de personne et d'organisation. Cependant, ANERsys était meilleur que CasANER dans la reconnaissance des noms de lieu. Dans ce qui suit, nous discutons les raisons de ce résultat obtenu. Par ailleurs, nous montrons également les avantages rajoutés par CasANER dans le corpus ANERcorp.

6.4. Problèmes rencontrés lors de l'application de CasANER sur ANERCorp

En réalité, il existe certains cas non détectés par CasANER comme les abréviations. En fait, nous n'avons pas traité les abréviations en élaborant notre système. En outre, CasANER ne trouvait aucune solution pour ces abréviations qui apparaissent avec une fréquence importante surtout dans les noms d'organisation comme <ORG> سي أن أن </ORG> ou <ORG> في تي جي </ORG>. De plus, CasANER reconnaît toujours les noms d'organisation à travers des noms d'organisation. Dans ce cas, les fameuses organisations non précédées par un indicateur ne seront pas reconnues. En effet, c'est le cas du nom d'organisation « الفيفا/ FIFA » (<ORG> الفيفا </ORG>) et pleins d'autres noms. Il faut dire aussi que CasANER ne reposait pas sur un dictionnaire qui stocke spécialement les noms d'organisation. Il utilisait que des règles d'extractions qui sont moins sûres que les entrées connues et fixes dans une liste comme celle utilisée par ANERsys.

Un autre cas non abordé par CasANER est celui des d'ENA écrites en d'autres langues. Par exemple, si une ENA française traduit alors notre système de REN va faire de son mieux pour la reconnaître sinon si elle est écrite en français, CasANER est dédié uniquement à la langue arabe. Toutefois, il existe plusieurs ENA écrites en d'autres langues dans ANERcorp et elles étaient reconnues par ANERsys comme par exemple <PERS> Charles I </PERS> et <ORG>

El Telegramma Del Rif </ORG>. Après l'analyse d'ANERcorp annoté par ANERsys, il faut signaler une erreur très remarquable. En fait, il existe plusieurs ENA reconnues et annotées par erreur comme par exemple « السودان/ Soudan » qui a été reconnue en tant qu'une ENA de la catégorie nom de personne.

6.5. Enrichissement de ANERcorp grâce au système CasANER

CasANER a pu contribuer à enrichir ANERcorp. Cet enrichissement concerne en premier lieu le raffinement fourni grâce à la typologie étendue sur laquelle CasANER repose. Cette typologie englobe une diversité de catégories ayant un niveau de granularité important. De plus, le niveau de raffinement a touché le quatrième niveau dans certaines catégories comme par exemple la catégorie Nom de lieu. En second lieu, CasANER a pu normaliser l'annotation des ENA existantes dans ANERcorp à travers l'application de la cascade de synthèse. Notre système a pu résoudre le fameux phénomène linguistique qui est l'agglutination qui n'était pas traité par ANERsys. Par exemple, l'ENA « وقاسم العزام / Et Kacim Al-Aazam » pose un problème d'agglutination vu qu'il existe une conjonction attachée au prénom. Malheureusement, ANERsys l'annote comme suit : <PERS>وقاسم العزام </PERS>. Néanmoins, CasANER cherche à protéger la conjonction par une balise symbolique pour avoir la forme de <Prps> و </Prps> suivie par l'ENA : <PERS> <forename> قاسم </forename> <surname> العزام </surname> </PERS>.

7. Conclusion

Dans ce chapitre, nous avons expérimenté les résultats obtenus par les travaux déjà implémentés dans le chapitre précédent. Cette expérimentation a facilité le passage à la phase de l'évaluation de nos deux systèmes CasANER et ASRextractor implémentés sous la plateforme linguistique Unitex. L'évaluation est faite à la base de mesures de performance qui ont montré que nos résultats sont encourageants. Pour étudier les lacunes de chaque système et localiser les points à améliorer, nous avons effectué une évaluation par catégorie d'ENA pour le système CasANER et une évaluation par type de RS pour le système ASRextractor. Nous avons signalé quelques erreurs rencontrées qui ont baissé la valeur de rappel pour les deux systèmes. Ces erreurs peuvent être dues au manque de contraintes suffisantes dans les règles d'extraction qui sont créées manuellement. De plus, la performance de ces règles d'extraction est liée également à la couverture des dictionnaires exploités qui nécessitent une amélioration. Pour les résultats du système ASRextractor, nous avons trouvé des anomalies à cause du système CasANER à savoir l'existence des ENA non reconnues ou encore reconnues en erreur. Mais, les deux systèmes ont souffert de la structure de la Wikipédia qui comporte des fautes de frappe.

Le système CasANER a subi un processus de comparaison avec un système de REN appelé ANERsys à travers son application sur le corpus ANERcorp. Notre système a excellé au niveau de la reconnaissance des noms de personne et nom d'organisation. Nous avons pu détecter quelques erreurs commises par l'ANERsys et par CasANER. Néanmoins, nous avons trouvé que le système CasANER a offert une représentation normalisée au corpus ANERcorp. De plus, notre système a réussi à résoudre le problème d'agglutination qui n'était pas traité par le système ANERsys. La comparaison de deux systèmes offre l'opportunité de tester la performance du système élaboré. Pour cette raison, nous tentons dans des futurs travaux à comparer ASRextractor avec d'autres systèmes d'extraction des RS entre les ENA.

Conclusion générale

Dans le présent travail, nous avons pu réaliser un système de REN que nous avons appelé CasANER basé sur une cascade de transducteurs. Ce système permet de reconnaître les ENA selon une typologie établie à la suite d'une étude linguistique profonde. Il exploite le niveau de granularité de la typologie déjà mentionné pour annoter rigoureusement les composantes des ENA reconnues. Le système CasANER repose sur diverses natures de transducteurs, analyse, filtrage et généralisation d'étiquetage, agissant sur un corpus extrait à partir de la Wikipédia arabe. De plus, il a permis de traiter le phénomène d'agglutination et d'effectuer une analyse syntagmatique pour les ENA imbriquées. Le système CasANER se base sur des dictionnaires de haute couverture qui ont participé à leur tour de guider le processus de reconnaissance. Par ailleurs, il s'appuie sur des règles de productions fortes ainsi que des grammaires regroupant les indicateurs classifiés dédiés à déclencher les ENA.

Par la suite, nous avons pu élaborer un système d'extraction des RS entre les ENA appelées ASRextractor basé sur une cascade de transducteurs. Le système ASRextractor agit sur un corpus Wikipédia arabe qui est le fruit d'un processus de REN effectué par le système CasANER. ASRextractor est un extracteur de RS selon une typologie de types identifiés lors d'une exploration et d'une analyse du corpus déjà annoté. Ce système traite les ENA qui possèdent une catégorie d'appartenance et celles qui ne sont pas catégorisées par le système CasANER. Cependant, les RS extraites sont annotées selon la norme TEI pour fournir des balises compréhensibles et structurées. Le système ASRextractor se base sur un dictionnaire sémantique prioritaire stockant les indicateurs sémantiques collectés et associés à des traits. Ce système se base aussi sur des fonctionnalités avancées lors de la création des transducteurs. Pour cette raison, il a pu résoudre l'éloignement des chemins entre les ENA reliant une RS.

Puis, nous avons créé des post-traitements sous forme de modules qui se basent sur des cascades de transducteurs. Le premier module est dédié à la normalisation pour transformer l'annotation des ENA reconnues par le système CasANER selon la norme TEI. Ce module de normalisation agit sur le corpus annoté, plus précisément la version XML après l'application du CasANER. Le deuxième module traite également les ENA qui se situent au sein des RS extraites par le système ASRextractor pour récupérer la forme brute. Ce module fonctionne sur la version issue directement d'ASRextractor. Nous avons pu améliorer le module arabe sous la plateforme Unitex qui nous a permis de réaliser tous nos travaux. Cette amélioration s'effectue à travers l'augmentation de la couverture des dictionnaires existants et la création de nouveaux dictionnaires à partir d'une banque d'arbres syntaxiques. D'ailleurs, cette création se base aussi sur une cascade de transducteurs.

L'implémentation de toutes les cascades de transducteurs élaborées est faite grâce à la plateforme Unitex. En fait, le regroupement de tous les transducteurs est assuré par l'outil CasSys. Cet outil a participé en grande partie à garantir un bon enchaînement de fonctionnement de nos systèmes et les modules implémentés. Le système CasANER englobe au total 178 graphes tandis que le système ASRextractor contient 199 graphes. Le module de normalisation et celui de récupération de la forme brute des ENA au sein des RS comportent chacun 2 graphes. L'élaboration des graphes se fait via de nouvelles techniques appliquées sur les graphes dans Unitex à savoir le mode morphologique et la notion des variables, etc.

Pour ce faire, nous avons effectué un état de l'art qui se compose de deux parties. Dans la première partie, nous avons étudié l'extraction des EN qui constitue une initiation à l'extraction des RS entre elles. Cette étude nous a permis de découvrir les définitions proposées dans la littérature pour délimiter les EN. Ces définitions varient selon la nature de corpus analysé. De plus, nous avons distingué les travaux de catégorisations effectuées depuis les conférences MUC qui visaient à améliorer le champ de catégorisation. Nous avons discuté les travaux antérieurs pour la tâche de REN et celle d'extraction des RS. Cette discussion nous a aidé à explorer les formalismes et les techniques utilisés et les résultats obtenus pour choisir la stratégie adéquate afin de réaliser nos systèmes. De plus, nous avons compris les ressources libres plus précisément la Wikipédia et quelques travaux qui l'exploitent pour effectuer des processus d'enrichissement. La présentation des normes d'annotation nous a permis d'étudier les domaines d'applications de la norme TEI et de voir son efficacité.

Dans la deuxième partie de l'état de l'art, nous avons étudié les automates et les transducteurs. Nous avons pu discerner l'importance des transducteurs à travers l'illustration des applications qui les exploitent. En fait, les transducteurs ont montré leur fiabilité quand ils sont regroupés au sein d'une cascade avec un ordre spécifique. Cette partie de l'état de l'art a facilité la compréhension de la plateforme Unitex et ses fonctionnalités avancées comme la création de graphes avec différents types de boîtes. En plus, Unitex intègre l'outil CasSys pour créer les cascades et choisir leurs modes de passage. Ces modes ont permis de générer des sorties répondant aux besoins des utilisateurs.

Nous avons effectué aussi une étude linguistique qui se compose également de deux parties. La première étude linguistique est basée sur le corpus d'étude extrait à partir de la Wikipédia arabe. La richesse et la diversité de styles d'écriture dans le corpus exploité nous a accordé l'opportunité d'effectuer un processus de catégorisation très étendue de sorte que nous avons proposé une typologie d'ENA large et détaillé englobant toutes les catégories et sous-catégories identifiées. Cette typologie fait appel à tous ses sous-typologies pour la catégorie Nom de lieu.

En fait, c'est une épreuve montrant que le corpus exploité favorise l'évaluation des ENA. Notamment, diverses formes d'ENA pouvant être assignées aux catégories et sous catégories détectées ont été identifiées en tenant compte les différentes cultures entre différents pays arabes. Il faut mentionner que l'établissement de la typologie d'ENA a été basée sur une définition claire et précise que nous avons proposée.

De plus, nous avons fait une deuxième étude linguistique basée sur la version annotée du corpus Wikipédia arabe après le passage du système CasANER. Nous avons pu profiter du raffinement de la typologie d'ENA et du bon niveau de granularité qu'elle offre pour identifier une typologie des types de RS entre les ENA. De même, cette typologie s'appuie également sur une définition de RS concrète et descriptive et englobe dix-huit types. Cette typologie contient aussi les RS qui relient des ENA non catégorisées. Par ailleurs, certaines RS possèdent des sous types grâces à l'annotation détaillée des ENA.

L'évaluation de nos systèmes a été faite via les mesures de performance. Cette évaluation montre que le système CasANER excelle en particulier dans la reconnaissance de la catégorie évènement. Ce système a donné un bon rendement aussi dans la reconnaissance des noms de personne car cette catégorie est bien étudiée à travers une grammaire de mots déclencheurs et trois ensembles de transducteurs d'analyse, de filtrage et de généralisation d'étiquetage. Les noms de lieu sont bien reconnus par le système CasANER car ils sont classifiés en des sous-catégories qui sont raffinées à leur tour. Le système ASRextractor excelle en quatre types de RS qui sont date de décès, équivalence date, nationalité et origine, car ils possèdent des règles de productions exhaustives. Ces cinq types de RS apparaissent dans des textes ayant une structure en commun. Notamment, ils sont toujours exprimés par des segments courts ce qui rend leur extraction plus facile. Le reste des types sont estimables puisque leur évaluation montre de bonnes valeurs. Pour le système CasANER, nous avons effectué une comparaison de sa performance avec celle d'un système appelé ANERsys. Ce dernier est basé sur l'approche statistique et équipé par un corpus appelé ANERcorp construit par l'auteur du système déjà mentionné. Pour passer notre système CasANER, nous avons créé un module de préparation de corpus ANERcorp afin d'effectuer la comparaison. Ce module se base également sur la cascade de transducteurs. En fait, le système CasANER a montré qu'il est capable de reconnaître les ENA ayant la catégorie nom de personne et nom d'organisation mieux que le système ANERsys qui a pu reconnaître les noms de lieux mieux que le nôtre. Enfin, le système a pu obtenir de bons résultats malgré qu'il n'ait pas exploité les dictionnaires associés au système ANERsys et qu'il n'ait utilisé que ceux propres à Unitex.

Dans nos futurs travaux, nous comptons élargir la taille de ce corpus à partir de la Wikipédia arabe. En outre, l'ajout de nouveaux articles offre l'opportunité de manipuler de nouveaux domaines et de découvrir de nouvelles ENA. Cela favorise l'obtention de nouvelles catégories déclenchant la présence des nouvelles RS. De plus, l'amélioration de la taille du corpus nous pousse à augmenter la couverture des dictionnaires exploités pour nos systèmes. Nous allons penser à tester d'autres types de corpus pour généraliser les systèmes élaborés à savoir les corpus journalistiques ou les rapports médicaux. Pour la construction automatique de dictionnaires, nous allons explorer de nouvelles ressources linguistiques pour les appliquer sur l'extracteur automatique que nous avons créé.

Nous allons étendre les fonctionnalités de notre système CasANER pour explorer de nouvelles catégories d'ENA afin de développer la typologie établie. Nous allons essayer de raffiner ces catégories pour faciliter la découverte des nouvelles RS à extraire par le système ASRextractor. Ensuite, nous allons exploiter les RS extraites et annotées pour la recherche des liens externes reliant les ENA à des ressources libres. Autrement, nous exploitons les différents types de RS déjà identifiées afin de trouver un lien reliant une ENA à d'autres ressources extérieures comme DBpédia. Nous allons relier également les ENA géographiques avec des ressources de localisation comme Géonames. Dans ce cas, les ENA auront non seulement des RS mais également des URI (Uniform Resource Identifier) pour stimuler l'interconnexion. Cette interconnexion permet d'assurer une meilleure interopérabilité des données sur le web et un accès sous une licence libre.

Les systèmes CasANER et ASRextractor génèrent des corpus semi-structurés contenant des ENA et des RS entre elles, reconnues et annotées selon la norme TEI. Dans cette optique, nous comptons fusionner ces deux systèmes afin d'avoir un nouveau système. Cela va nous permettre de faciliter leur application et d'aller vers l'approche hybride en exploitant les corpus annotés comme une base d'apprentissage.

Finalement, toutes les tâches déjà mentionnées vont participer à un projet de création d'un dictionnaire électronique d'ENA ayant une interface conviviale avec un champ de recherche augmenté. De plus, ce dictionnaire électronique comporte des ENA normalisées et bien définies via les relations entre elles et vers des ressources libres. Cela permettra, d'une part, d'avoir un dictionnaire extensible et capable de s'intégrer dans le Web d'une autre part. L'intégration de ce genre de ressource dans le web évite l'isolement de données et construit aussi un réseau d'information important pour la langue arabe.

Bibliographies

AbdelRahman S., Elarnaoty M., Magdy M. et Fahmy A. 2010. Integrated machine learning techniques for Arabic named entity recognition. International Journal of Computer Science Issues (IJCSI). Pp 27–36.

Abd El-Salam M. Sh., El Houby M.F. E., Al sammak A.K. et El-shishtawy T.A. 2016. Extracting Arabic Relations from the Web. International Journal of Computer Science & Information Technology (IJCSIT) Vol 8, No 1. Pp 85-102.

Abney, S. 1996. Partial Parsing via Finite-State Cascades. In Proc. of Workshop on Robust Parsing, 8th European Summer School in Logic, Language and Information, Prague, Czech Republic. Pp 8-15.

Aboaoga M. et Aziz MJA. 2013. Arabic person names recognition by using a rule based appraoch. Journal of Computer Science. Vol 9. Pp 922-927.

Abuleil S. 2006. Hybrid System for Extracting and Classifying Arabic Proper Names. Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17. Pp 205-210.

Aliane H., Guendouzi A., Mokrani A. 2013. Annotating Events, Time and Place Expressions in Arabic Texts. Proceedings of Recent Advances in Natural Language Processing. Pp 25–31, Hissar, Bulgaria.

Arnulphy, B., et Tannier, X. 2013. Entités Nommées Événement : guide d'annotation. Notes ET Documents LMSI. 35 pages.

Balvet A. 2002. LIZARD, un assistant pour le développement de ressources linguistiques à base de cascades de transducteurs. Nancy, RÉCITAL. Pp 24-27.

Basile, P., Caputo, A., Gentile, A. L., & Rizzo, G. 2016. Overview of the EVALITA 2016 Named Entity Recognition and Linking in Italian Tweets (NEEL-IT) Task. In CLiC-it/EVALITA.

Bates M. 1978. The theory and practice of augmented transition network grammars. In Natural language communication with computers. Springer, Berlin. Pp 191-254.

Belainine B. 2017. Classification supervise de textes courts et bruités : Application au domaine des médias sociaux. Mémoire de maîtrise en Informatique, université de Québec à Montréal. 113 pages.

Ben Abacha A. et Zweigenbaum P. 2011. Automatic Extraction of Semantic Relations between Medical Entities: A Rule Based Approach. Journal of Biomedical Semantics 2. Suppl 5: S4. PMC. Web.

Béchet F. et Charton E. 2010. Unsupervised knowledge acquisition for extracting named entities from speech. In Acoustics Speech and Signal Processing (ICASSP), IEEE International Conference. Pp 5338-5341.

Ben Ismail S., Maraoui H., Haddar K., Romary L. 2017. ALIF editor for generating Arabic normalized lexicons. The International Conference on Information and Communication Systems (ICICS 2017), Irbid, Jordan. IEEE. Pp 70-75.

Ben Mesmia F. 2013. Enrichissement de Prolexbase par des liens vers des ressources libres. Mémoire de mastère à la Faculté des Sciences de Sfax (FSS). Soutenu le 23/12/2013. 121 pages.

Benajiba Y. et Rosso P. 2008. Arabic Named Entity Recognition using Conditional Random Fields. In Proceedings of Workshop on HLT and NLP within the Arabic World, LREC. Pp. 143-153.

Borthwick A., Sterling J., Agichtein E. et Grishman R. 1998. NYU: Description of the MENE named entity system as used in MUC-7. In 7th Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia.

Bikel D. M., Miller S., Schwartz R. et Weischedel R. 1997, March. Nymble: a high-performance learning name-finder. In Proceedings of the fifth conference on Applied natural language processing. Association for Computational Linguistics. Pp. 194-201.

Ben Hamadou A., Piton O., Fehri H. 2011. Multilingual Extraction of functional relations between Arabic Named Entities using NooJ platform. International Conference and Workshop, Komotini, Greece. Proceedings of the Nooj 2010 International Conference and Workshop. Pp 192-202.

Biltawi, M., Awajan, A., Tedmori, S., et Al-Kouz, A. 2016. Exploiting Multilingual Wikipedia to improve Arabic Named Entity Resources. International Arab Conference on Information Technology (ACIT). Sultan Moulay Slimane University, Beni Mellal, Morocco. Pp 29-32.

Boujelben I., Jamoussi S., et Hamadou A. B. 2014. A hybrid method for extracting relations between Arabic named entities. Journal of King Saud University-Computer and Information Sciences, 26(4). Pp 425-440.

Btoush M-H., Alarabeyyat A. and Olab I. 2016. Rule Based Approach for Arabic Part of Speech Tagging and Name Entity Recognition. International Journal of Advanced Computer Science and Applications (IJACSA). Vol. 7. No. 6. Pp 331-335.

Chinchor N. 1998. Overview of MUC-7/MET-2. In Proceedings of the Seventh Message Understanding Conference (MUC-7). Pp 1–4. Fairfax, VA, USA.

Ciravegna, F., Lavelli, A. 1999. Full text parsing using cascades of rules: An information extraction perspective. In Proceedings of EACL'99, Bergen, Norway. Pp 102-109.

Darwish K. et Gao W. 2014. Simple Effective Microblog Named Entity Recognition: Arabic as an Example. In LREC. Pp. 2513-2517.

Doumi, N., Lehireche, A., Maurel, D., Ali Cherif, M. 2013. La conception d'un jeu de ressources libres pour le TAL arabe sous Unitex. Paper presented at the TRADETAL2013, Colloque international en Traductologie et TAL, Oran - Algeria, 5-6 may. Pp. 5-6.

Doumi, N., Lehireche, A., Maurel, D., & Khater, M. 2015. Using finite-state transducers to build lexical resources for Unitex Arabic package. In ACTES DU COLLOQUE CECTAL. Pp. 83-93.

Doddington G., Mitchell A, Przybocki M., Ramshaw L., Strassel S., et Weischedel R. 2004. The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation. In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC). Pp 837–840. Lisbon, Portugal.

Dufournaud N., Demonet M. L., Uetani T., Vincent T., le Rolle V., Bontemps L., ... & Jimenes R. 2012. Manuel d'encodage TEI-Renaissance et temps modernes. Version 3. 69 pages.

Ehrmann M. 2008. Les Entités Nommées, de la Linguistique au TAL - Statut Théorique et Méthodes de Désambiguïsation. Thèse de doctorat, Université Paris 7 Denis Diderot. Soutenue le 2 Juin 2008, 295 pages.

Ciravegna F., Lavelli A., Mann N., Gilardoni L., Mazza S., Ferraro M., Matiasek J., Black W.J, Rinatdi F. et Mowatt D. 1999. FACILE: Classifying texts integrating pattern matching and information extraction. In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence. Sweden. Pp 890-897

Friburger N. 2002. Reconnaissance automatique des noms propres ; application à la classification automatique de textes journalistiques. Thèse de doctorat, Université François-Rabelais de Tours. Soutenue le 2 décembre 2002. 172 pages.

Friburger N., Maurel D. (2004), Finite-state transducer cascade to extract named entities in texts, Theoretical Computer Science. Vol 313. Pp 94-104.

Eshkol-Taravella, I. 2015. La définition des annotations linguistiques selon les corpus : de l'écrit journalistique à l'oral. Mémoire d'habilitation à Diriger des Recherches. Université d'Orléans. 199 pages.

Ezzat M. 2014. Acquisition de relations entre entités nommées à partir de corpus. Ordinateur et société [cs.CY]. Institut National des Langues et Civilisations Orientales- INALCO PARIS - LANGUES O'.

Ghamnia, A. 2016. Extraction de relations d'hyperonymie à partir de Wikipédia. Actes de la conférence conjointe JEP-TALN-RECITAL. Vol 3. Pp 40-51.

Gravier, J.F. Bonastre, S. Galliano, E. Geoffrois, K. Mc Tait and K. Choukri. 2004. ESTER, une campagne d'évaluation des systèmes d'indexation d'émissions radiophoniques. Proc. Journées d'Etude sur la Parole.

Grishman R et Sundheim B. 1996. Message Understanding Conference- 6: A Brief History. Dans Proceedings of the 16th conference on Computational linguistics (COLING'96). Pp 466–471, Copenhagen, Denmark.

Grouin C., Galibert O., Rosset S., Quintard L. et Zweigenba P. 2011. Mesures d'évaluation pour entités nommées structurées. Dans Évaluation des méthodes d'Extraction de Connaissances dans les Données, Brest, France.

Habash N., Rambow O. et Roth R. 2009. MADA+ TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In: Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt. Pp. 62.

Hammouda, N. G. et Haddar, K. 2016. Integration of a segmentation tool for Arabic corpora in NooJ platform to build an automatic annotation tool. In International NooJ Conference. Springer, Cham. Pp 89-100.

Hammouda, N. G. et Haddar, K. 2017. Parsing Arabic Nominal Sentences with Transducers to Annotate Corpora. Computación y Sistemas, Vol 21, n° 4. Pp 647-656.

Hobbs J., Appelt D., Bear J., Israel D., Kameyama M., Stickel M., Tyson M. 1997. FASTUS: A cascaded finite-state transducer for extracting information from natural-language text. Finite-State Language Processing, MIT Press Pp 383-406.

Hkiri E., Mallat S., Zrigui M. 2016. Système hybride pour la reconnaissance des entités nommées arabes à base des CRF. Actes de la conférence conjointe JEP-TALN-RECITAL 2016, volume 2 : TALN. Pp 539-546.

Isozaki H. 2001. Japanese named entity recognition based on a simple rule generator and decision tree learning. In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics. Pp 314-321.

Kanya N. and Ravi T. 2016. Named Entity Recognition From Biomedical Text –An Information Extraction Task. ICTACT Journal on Sort Computing. Vol 6. Issue 04. Pp 1302-1307.

Kobilarov G. et al. 2009. Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make. In: European Semantic Web Conference. Springer, Berlin, Heidelberg, 2009. Pp 723-737.

Kostov, J. 2016. FlexiMac 1.1.-Conjugueur automatique des verbes macédoniens. In Conférence conjointe JEP-TALN-RECITAL. Pp 18-20.

Küçük D. et Yazıcı, A. 2012. A Hybrid Named Entity Recognizer for Turkish. Expert Systems with Applications. Volume 39. Pp 2733–2742.

Kurdi M. Z. 2018. Traitement automatique des langues et linguistique informatique 2. Volume 2. ISTE Editions. 280 pages.

Lafourcade M. et Ramadier L. 2016. Semantic Relation Extraction with Semantic Patterns: Experiment on Radiology Report. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC). Pp 4578–4582. Portorož Slovenia.

Landomiel, F., Gupta, A., Maurel, D., Poupon, A. 2017. Préliminaire à la construction d'un réseau de signalisation en biologie systémique. Atelier sur la Fouille de Textes. Pp 31-42.

Lahbib W., Bounhas I., Elayeb B., Evrard F. and Slimiani Y. 2013. A Hybrid Approach for Arabic Semantic Relation between Semantic Relation Extraction. Proceedings of the twenty-sixth International Florida Artificial Intelligence Research Society Conference. Pp 315–320.

Lhioui M., Haddar H., Romary L. 2016. A new method for interoperability between lexical resources using MDA approach. 2nd International Conference on Advanced Intelligent Systems and Informatics (AISI). Cairo, Egypt. Pp 64-74.

Lison P. 2004. Grammaires Locales : Principes, Modélisation et Utilisation1. "Introduction au Traitement du Langage Naturel" (FLTR 2620). Université Catholique de Louvain, Faculté de Philosophie et Lettres, Centre de Traitement Automatique du Langage. 11 pages.

Maurel D., Friburger N., Antoine J.-Y., Eshkol-Taravella I., Nouvel D. 2011. Cascades de transducteurs autour de la reconnaissance des entités nommées. Traitement automatique des langues, 52(1). Pp 69-96.

Maurel D., Friburger N., Eshkol I. 2009. Who are you, you who speak? Transducer cascades for information retrieval. In Proceedings of 4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznan, Poland. Pp. 220-223.

Maurel D., Friburger N, Eshkol I., Antoine J-Y. 2013. Explorer des corpus à l'aide de CasSys. Application au Corpus d'Orléans. TEXTE ET CORPUS, N°4 Actes des sixièmes Journées de la Linguistique de Corpus. Pp 189-195.

Maraoui H. et Haddar K. 2015. Automatisation de l'encodage des lexiques arabes en TEI. Colloque pour les Étudiants Chercheurs en Traitement Automatique du Langage naturel et ses applications. Pp 74-82.

Maraoui H., Haddar K. and Romary L.2018: Segmentation tool for hadith corpus to generate TEI encoding. AISI'19 Conference, Springer. Volume n°845. Pp 252-260.

Merchant, R., Okurowski, M. et Chinchor, N. 1996. The multilingual entity task (MET) overview. In Proceedings of a workshop on held at Vienna, Virginia. Pp 445–447, Morristown, NJ, USA. Association for Computational Linguistics.

Mesfar S. 2007. Named entity recognition for Arabic using syntactic grammars. In: Proceedings of the 12th international conference on application of natural language to information systems. Berlin: Springer. Pp 305-316.

Mesfar, S. 2008. Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard. Thèse de doctorat. Besançon, France : Université de Franche Comté. Soutenu le 24 novembre 2008.

Mohammed NF and Omar N. 2012. Arabic named entity recognition using artificial neural network. Journal of Computer Science. Pp 1285-1293.

Mokrane, A., Antoine, J. Y., & Friburger, N. 2008. Cascades de transducteurs pour le chunking de la parole conversationnelle : l'utilisation de la plateforme CasSys dans le projet EPAC. In Procceedings of the 15ème Conférence sur le Traitement Automatique du Langage naturel (TALN).

Mikheev Andrei, Grover Claire, et Moens Marc. 1998. Description of the LTG system used for MUC- 7. Dans Proceedings of the Seventh Message Understanding Conference (MUC-7). Pp 1– 12, Fairfax, VA, USA.

Nadeau N., Sekine S. 2009. A survey of named entity recognition and classification», Satoshi Sekine and Elisabete Ranchhod, ed., John Benjamins publishing company. Pp 3-28.

Nadeau N., Sekine S. 2007. A survey of named entity recognition and classification. Lingvisticae Investigations. Vol 30. Pp 3-26.

Oudah M. et Shaalan K. 2012. A pipeline Arabic named entity recognition using a hybrid approach. Dans Proceedings of the 24th International Conference on Computational Linguistics (COLING'12). Pp 2159–2176.

Oudah, M., et Shaalan, K. 2017. NERA 2.0: Improving coverage and performance of rule-based named entity recognition for Arabic. Natural Language Engineering, 23(3), pp 441-472.

Paumier S. 2003. De la Reconnaissance de Formes Linguistiques à l'Analyse Syntaxique, Thèse de Doctorat en Informatique, Université de Marne-la-Vallée soutenue le 4 juillet 2003. 198 pages.

Paumier S. 2017. UNITEX 3.2 ALPHA MANUEL D'UTILISATION. Université Paris-Est Marne-la-Vallée. 394 pages. Date de cette version : 28 septembre 2017.

Poibeau T. 2003. Extraction automatique d'information, du texte brut au web sémantique, Lavoisier. ISBN 2-7462-0610-2.

Poibeau T. 2003b. The Multilingual Named Entity Recognition Framework. La conférence internationale EACL 2003, pages 155-158, 2003.

Poibeau T. 2011. Traitement automatique du contenu textuel. Editeur : Paris, Hermès Science publications, ISBN : 9782746231917 2746231913, 222 pages.

Paik W., Liddy E. D., Yu E., et McKenna M .1996. Categorizing and standardizing proper nouns for efficient information retrieval. Publié dans le livre Corpus processing for lexical acquisition. Pp 61–73.

Ramesh D. and Sanampudi S-K. 2016. A Hybrid model for Named Entity Recognition in Biomedical text. International Journal of Scientific & Engineering Research. Volume 7, Issue 6. ISSN 2229-5518. pp 1164- 1166.

Saleh I., Tounsi L. and Van Genabith J. 2011. ZamAn and Raqm: Extracting Temporal and Numerical Expressions. In Arabic in Information Retrieval, Lecture notes in computer science vol. 7097. Pp 562-573.

Salleh, M. S., Asmai, S. A., Basiron, H., et Ahmad, S. 2017. A Malay named entity recognition using conditional random fields. In 5th International Conference of Information and Communication Technology (ICoIC7). Pp 1-6.

Satoshi S. et HitoshiI. 1999. IREX project overview. In: Proceedings of the IREX Workshop. Pp 7-12.

Sauri R., Knippen R., Verhagen M., Pustejovsky J. 2005. Evita: A Robust Event Recognizer For QA Systems. In HLT'05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Pp 700-707, Morriston, NJ, USA : Association for Computational Linguistics.

Sellami, R., Sadat, F., et Belguith Hadrich, L. 2012. Extraction de lexiques bilingues à partir de Wikipédia. Atelier de Traitement Automatique des Langues Africaines, JEP-TALNRECITAL. Pp. 107-117.

Sellami, R., Sadat, F., et Belguith, L. H. 2013. Traduction automatique statistique à partir de corpus comparables : application aux couples de langues arabe-français. In CORIA. Pp 431-440.

Semmar N., Zennaki O., Laib M. 2016. Etude de l'impact d'un lexique bilingue spécialisé sur la performance d'un moteur de traduction à base d'exemples. Actes de la conférence conjointe JEP-TALN-RECITAL. Vol 2. TALN. Pp 84 – 97.

Serrano L., Charnois T., Brunessaux S., Grilheres B., Bouzid M. 2012. Combinaison d'approches pour l'extraction automatique d'événements. TALN'2012. Vol 2. Pp 423– 430. Grenoble, France.

Silberztein M., Poibeau T., Balvet A. 2001. Tutoriel : Intex et ses applications informatiques, Actes de la huitième conférence sur le Traitement Automatique des Langues Naturelles. Pp.145-174, Tours.

Silberztein, M. 2002. Manuel de NooJ (2003). Télécharger à partir de http: // www. nooj4nlp. net. 394 pages.

Shaalan K and Raza H. 2007. Person named entity recognition for Arabic. In: Proceedings of the 5th workshop on important unresolved matters. p. 17-24.

Shaalan K. 2010. Rule-based approach in Arabic natural language processing. The international Journal of Information and Communication Technologies (IJITC). Vol 3. Pp 11-19.

Shaalan K. et Raza H. 2009. NERA: Named entity recognition for Arabic. Journal of the American Society for Information Science and Technology, 60(9). Pp1652–1663.

Shaalan K. et Oudah M. 2014. A hybrid approach to Arabic named entity recognition. Journal of Information Science. Vol 40(1). Pp 67-87.

Sharma P., Sharma U. and Kalita J. 2016a. Named Entity Recognition in Assamese: A hybrid approach. International Conference on Advances in Computing, Communications and Informatics (ICACCI-2016), Jaipur, India. Pp 2114-2120.

Sharma P., Sharma U. and Kalita J. 2016b. Named Entity Recognition in Assamese. Journal of Computer Applications, Volume 142 - No.8. Pp 1-8.

Soualah M.O. et Hassoun M. 2012. A TEI P5 Manuscript Description Adaptation for Cataloguing Digitized Arabic Manuscripts. Journal of the Text Encoding Initiative. Issue 2. Lien du papier: https://journals.openedition.org/jtei/398.

Straubing H. et Weil P. 2012. An Introduction to Finite Automata and their Connection to Logic. Modern applications of automata theory. Pp 3-43.

Talha M. Boulaknadel S. et Aboutajdine D. 2014. RENAM : Système de Reconnaissance des Entités Nommées Amazighes. 21ème Traitement Automatique des Langues Naturelles, Marseille. Vol. 2. Pp. 517-524

Text Encoding Initiative Consortium. 2018. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Edited by C.M. Sperberg-McQueen and Lou Burnard for the ACH-ALL-CACL. Version 3.2.0. 1887 pages. Dernière mise à jour le 31 janvier 2018.

Tourille J., Ferret O., Tannier X., Névéol A. 2017. Temporal information extraction from clinical text. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Pp 739–745, Valencia, Spain.

Tjong Kim Sang, E. F., et De Meulder, F. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. Dans les actes de CoNLL-2003. Pp 142-147.

Woods W. A. 1970. Transition network grammars for natural language analysis. Communications of the ACM. Volume 13. No. 10. Pp 591-606.

Yao L., Liu H., Liu Y., Li X., Anwar M-W. 2015. Biomedical Named Entity Recognition based on Deep Neutral Network. International Journal of Hybrid Information Technology Vol.8, No. 8. Pp 279-288.

Yegnanarayana, B. 1994. Artificial neural networks for pattern recognition. In Sadhana.Volume 19. Part 2. Pp 189-238.

Zaghouani W., Pouliquen B., Ebrahim M. et Steinberger R. 2010. Adapting a resource-light highly multilingual named entity recognition system to Arabic. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). Pp 563–567.

Zribi I., Hammami Mezghani S., Hadrich Belguith L. 2010 L'apport d'une approche hybride pour la reconnaissance des entités nommées en langue arabe. TALN. Pp 59.