Méthodes d'acquisition terminologique en arabe : Application au domaine médical

Wafa Neifar

NNT : 2019SACLS085

Thèse de doctorat

Méthodes d'acquisition terminologique en arabe : Application au domaine médical

Thèse de doctorat de l'université Paris-Saclay
préparée à L'université Paris-Sud

Ecole doctorale n°580 Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Orsay, le 18/03/2019, par

WAFA NEIFAR

Composition du Jury :

Nadia ESSOUSSI Professeur, Université de Tunis (LARODEC)	Président
Béatrice DAILLE Professeur, Université de Nantes (LINA)	Rapporteur
Frédéric BECHET Professeur, Aix Marseille Université (LIS)	Rapporteur
Olivier FERRET Ingénieur-Chercheur, CEA-LIST (LVIC)	Examinateur
Pierre ZWEIGENBAUM Directeur de recherche, CNRS (LIMSI)	Directeur de thèse
Lamia HADRICH BELGUITH Professeur, Université de Sfax (MIRACL)	Directeur de thèse
Thierry HAMON Maître de conférences, Université Paris-13 (LIMSI-CNRS)	Encadrant
Mariem ELLOUZE KHEMAKHEM Maître assistant, Université de Sfax (MIRACL)	Encadrant

Remerciements

Et vient le moment d'écrire les remerciements...

Quand je pense à toutes les personnes sans qui ce travail n'aurait probablement pas pu être achevé, je me dis que ce n'est pas pour rien que cette page précède la présentation de tout travail de thèse.

Je tiens tout d'abord à remercier Pierre Zweigenbaum, mon directeur de thèse en France, pour ses précieux conseils, ses encouragements, sa bonne humeur et son soutien moral. Malgré ses responsabilités, il a su être présent quand il le fallait.

Je remercie également, ma directrice de thèse en Tunisie, Lamia Hadrich Belguith, pour son encouragement, ses conseils et critiques constructives qui ont contribué à l'aboutissement de cette thèse.

Je voudrais aussi exprimer mes plus profonds remerciements à mon co-directeur en France, Thierry Hamon, pour sa confiance, pour la liberté qu'il m'a accordée et pour la qualité de son encadrement. Je voudrais aussi le remercier pour ses qualités personnelles et humaines qui ont aussi beaucoup contribué à la réalisation de ce travail.

En parlant de l'encadrement, je remercie Mariem Ellouze Khemakhem, ma co-directrice en Tunisie, pour ses conseils utiles. Ses conseils m'ont permis de prendre les bonnes décisions et mener à bien ce travail.

Je tiens à remercier mes deux rapporteurs de thèse Béatrice Daille et Frédéric Béchet pour avoir accepté de relire ce manuscrit, pour l'intérêt qu'ils ont porté à mon travail ainsi que les remarques et les suggestions qu'ils m'ont faites.

Je tiens aussi à remercier les examinateurs de ma thèse Olivier Ferret et Nadia Essoussi d'avoir accepté d'être parmi le jury de ma thèse. Merci beaucoup pour vos remarques et conseils.

La recherche scientifique c'est aussi faire partie d'une équipe, c'est pourquoi je suis particulièrement honorée d'avoir côtoyé les membres des deux laboratoires de recherche LIMSI et MIRACL que je remercie bien.

Un grand merci à mon mari Mohamed Ali Labidi pour sa compréhension, ses nombreux sacrifices, sa patience, son soutien moral, son grand amour et sa confiance en moi. Tu n'as pas cessé de m'encourager et de m'offrir des conditions favorables durant ces années d'études.

Je dédie ce travail à mon cher papa Mohamed pour lequel j'exprime mon amour et ma

gratitude pour ses sacrifices et son soutien moral, à mes frères Aymen et Rami, mes belles sœurs, mes beaux parents, tous les membres de ma famille et mes chers amis pour leur soutien moral et leurs encouragements durant la période de ma thèse.

Et pour finir, je ne remercierai jamais assez, et tous les mots du monde ne suffiront pas pour cela, celle qui a toujours été là, derrière moi, qui m'a toujours poussée et encouragée, soutenue dans les moments difficiles, qui a cru en moi, je pense bien sûr à ma mère Sabah, cette thèse est avant tout pour toi et à toi, alors merci du fond du cœur.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Sommaire

1.1 Motivations

Au cours des dernières décennies, la production textuelle et l'évolution des techniques d'analyse automatique de ces textes ont augmenté de façon exponentielle dans tous les domaines et notamment le domaine médical. C'est le cas dans la plupart des langues naturelles dont le nombre de locuteurs est important. De même, l'arabe standard moderne (MSA) fait partie de ces langues pour lesquelles il existe de plus en plus d'approches de Traitement Automatique des Langues (TAL).

Cependant, lorsqu'il s'agit de domaines de spécialité comme la médecine, le droit, l'informatique ou l'agriculture, ce constat est moins évident : le MSA n'est pas forcément la langue la plus utilisée dans tous les domaines de spécialité. Ainsi, si le droit, la géologie ou l'agriculture tendent à utiliser le MSA, en médecine, lors de la pratique et de l'enseignement, c'est la langue française ou anglaise qui est plus généralement utilisée par les spécialistes du domaine. Le français et l'anglais sont les langues utilisées dans les pratiques professionnelles dans le domaine de la santé dans les pays arabes où les documents, ordonnances, rapports, etc. sont principalement produits dans l'une de ces langues étrangères. De plus, la plupart des articles scientifiques sont publiés en anglais ou en français[Samy et al., 2012]. Cependant, la compréhension des notions spécialisées par le grand public, notamment les patients, est primordiale pour la santé publique Samy et al. [2012]. La constitution de terminologies dédiées au public d'un domaine ou ayant pour objectif de faciliter l'accès aux informations spécialisées et aux documents électroniques doit donc être favorisée.

Le travail de constitution de terminologie concerne des domaines de recherche divers et fait appel à de nombreuses disciplines comme la linguistique, la traduction, la science de l'information, ou le traitement automatique des langues. Ce travail est réalisé par des terminologues. Il vise à identifier les concepts d'une langue de spécialité et les termes qui les expriment, ainsi qu'à traiter la description, l'organisation et le transfert des connaissances [Sager, 1990]. La norme ISO 704 :2009(fr)décrit le travail terminologique en tant que tâche qui sert à la clarification et la normalisation des concepts et des terminologies afin de faciliter la communication entre humains. Des méthodes de TAL et d'acquisition terminologique automatique facilitent, elles, ce travail de constitution terminologique à partir de textes. Celles-ci doivent prendre en compte à la fois la pratique terminologique du domaine et les spécificités de chaque langue, étant donné la dimension linguistique du matériau textuel ISO 704 :2009(fr).

Dans le cadre de notre thèse, nous nous intéressons à l'acquisition terminologique automatique en tant que volet du travail terminologique. Notre objectif est de proposer différentes

méthodes afin de construire une terminologie monolingue ou bilingue pour l'arabe standard moderne. Nous appliquerons les méthodes proposées sur des textes issus du domaine médical. Pour cela, nous présentons dans cette introduction les deux éléments fondamentaux sur lesquels repose notre travail : la terminologie, la discipline à traiter et à automatiser, et l'arabe standard moderne, la langue principale de notre travail.

1.2 La terminologie

Dans sa première acception, la notion de terminologie désigne l'ensemble des expressions techniques, dites *termes*, exprimant les concepts d'une science, d'une technique ou d'un domaine spécifique particulier de l'activité humaine. Dans ce contexte, nous citons à titre d'exemple, le *Trésor de la Langue Française* (TLF) et sa version informatisée (TLFi) qui représente un dictionnaire des XIXe et XXe siècles en 16 volumes et 1 supplément. Il comporte 100 000 mots avec leur histoire, 270 000 définitions, 430 000 exemples et 350 millions de caractères. Le TLFi se distingue des autres dictionnaires électroniques existants par la finesse de la structuration des données.

C'est aussi l'étude scientifique des notions et des termes en usage dans les langues de spécialité ISO 1087 :1990. Aussi, du point de vue de l'usage, la terminologie présente un élément indispensable pour certaines activités qui nécessitent la représentation et le transfert des connaissances telles que la traduction technique, l'enseignement des langues, la rédaction technique, la documentation, l'ingénierie des langues, la normalisation technique, etc. Toute activité s'appuyant sur des connaissances spécialisées nécessite une terminologie [Cabré Castellví, 2003].

Il existe différents types de ressources terminologiques. Chacun d'entre eux est dédié à un usage particulier :

— Les base de données terminologiques multilingues, pour l'aide à la traduction ;

— Les thesaurus, pour les systèmes d'indexation automatique ;

— Les index structurés, pour les documentations techniques hypertextuelles ;

— Les référentiels terminologiques, pour les systèmes de gestion de données techniques ;

— Les glossaires de référence, pour les outils de communication interne et externe.

Dans cette section, nous résumons les aspects et les pratiques du travail du terminologue d'une manière générale mais aussi pour la langue arabe afin de mieux contraster les pratiques de construction terminologique en arabe par rapport à l'anglais et au français. Par la suite, comme

nous nous intéressons au domaine médical, nous mettons l'accent sur la construction terminologique de ce domaine. Puis, nous présentons les spécificités de la terminologie pour l'arabe standard moderne.

1.2.1 Pratiques terminologiques et usage

De nombreux chercheurs se sont intéressés à la construction et à la normalisation des terminologies. Les premières apparitions de la terminologie en tant que domaine de recherche datent de la fin du dix-neuvième siècle en Europe occidentale et en Russie [Rondeau, 1984 ; Rey, 1979]. Cependant, les premières bases de la terminologie en tant que discipline ont été posées au début du vingtième siècle par Eugène Wüster [Wüster, 1979]. Dans ses travaux, le terme est considéré comme l'étiquette d'un concept, désignant une notion de manière univoque, mono-référentielle et non contextuelle. Par la suite, cette discipline s'est développée et a évolué en créant des alliances avec d'autres disciplines comme l'informatique, la philosophie, la logique, la sociolinguistique, etc.

Ainsi, avec l'arrivée des outils automatiques et des corpus numérisés, même la théorie de la terminologie a évolué. Les terminologies sont désormais constituées d'un ensemble de termes adaptés aux besoins des usagers. Pour cela, chaque terme peut se voir associé des variantes. Tout dépend de l'univers sociodiscursif et de la structure sociale des personnes auxquelles la terminologie est destinée. La terminologie choisie doit être la plus proche possible du public visé afin de faciliter la compréhension des termes.

Par exemple, afin d'appliquer une approche socioterminologique au discours scientifique arabe du domaine du génie génétique, Abi Ghanem-Chadarevian [2016] montre que la circulation des termes est liée à différentes situations discursives. Plusieurs facteurs interviennent et influent sur la nature et la forme des termes utilisés lors de la construction d'une terminologie

- Le locuteur : étudiant, collègue, professeur, laborantin, chercheur, etc.
- L'interlocuteur : étudiant, collègue, professeur, laborantin, patient, non-expert, le proface, chercheur, etc.
- L'objectif du locuteur : convaincre, exposer des faits et des théories, avertir, éveiller la curiosité, sensibiliser, etc.
- Le lieu d'utilisation de la terminologie : laboratoire, discussion, cours, salle de conférence, publicité, à la télévision, etc.
- La forme du discours : texte écrit, communication orale, brochure, rapport médical, note de service, etc.

— L'espace géographique : pays européens, pays arabes, etc.

Dans le travail terminologique, il est important d'avoir un corpus de textes issus d'un domaine de spécialité et le représentant. Celui-ci permet aussi de mieux étudier et analyser l'usage et la circulation des termes dans une communauté linguistique et technique donnée. En effet, la particularité qui distingue un terme des autres unités porteuses de sens tient au fait que les termes appartiennent à un domaine spécifique. Nous pouvons alors définir un terme comme étant une unité linguistique qui exprime un concept de base dans un domaine de spécialité basé sur un vocabulaire et des usages linguistiques qui lui sont propres. Un terme peut correspondre à un seul mot ou à une seule unité lexicale, désigné comme un terme simple, ou être composé de plusieurs unités lexicales ou d'une unité phraséologique, alors appelé terme complexe.

Selon Wright and Budin [2001], la détermination et le choix des langues et des domaines d'application participent à la codification de la nature d'un terme. Autrement dit, le choix des termes à un seul mot, terme simple, ou à plusieurs mots, terme complexe, dépend des conventions spécifiques à la langue. Un terme peut apparaître comme le terme simple ***appendix*** (en anglais), ***l'appendice*** (en français) ou sous forme d'un terme complexe **الزائدة الدودية** (en arabe standard moderne).

Certains concepts et termes peuvent appartenir à des disciplines connexes ou des technologies convergentes. Aussi, un même terme peut exprimer différents concepts, contrairement aux préceptes de la Théorie Générale de la Terminologie [Wüster, 1979]. L'ambiguïté sera alors levée en précisant le domaine de spécialité [Pavel et al., 2002].

La construction des terminologies bilingues et multilingues constitue un champ de recherche à part entière. Elle consiste à identifier et repérer les termes et leurs traductions. Une terminologie bilingue n'est autre qu'un ensemble de deux terminologies unilingues corrélées [Gouadec, 1990]. Au-delà de ses connaissances linguistiques et de la terminologie du domaine de spécialité donné, le travail du terminologue doit alors se fonder sur une bonne connaissance des règles de formation lexicale dans la langue d'accueil et sur une bonne expérience en traduction et en rédaction technique. Ces connaissances représentent un précieux atout professionnel [Pavel et al., 2002]. Dagan et al. [1991] et Dyvik [1998] affirment que la constitution d'une terminologie dans différentes langues présente un atout à exploiter. Il s'agit de combler le manque terminologique dont souffrent d'autres langues, soit en enrichissant leurs vocabulaires par de nouveaux termes, soit en construisant à l'aide d'un graphe de traduction multilingue de nouvelles terminologies à partir de rien. La création et la maîtrise de ces terminologies multilingues serviront par la suite pour certaines pratiques comme la traduction spécialisée.

Le domaine médical n'échappe pas à ce phénomène. En effet, depuis la seconde moitié du vingtième siècle, plusieurs technologies multilingues ont été appliquées au domaine de la médecine. Notons par exemple les travaux de Sager et al. [1987] qui ont été fondateurs pour l'analyse automatique des textes. Par la suite, la nécessité du développement des réseaux internes dans les hôpitaux et les cliniques pour accéder aux informations médicales et les exploiter a motivé l'utilisation de ressources terminologiques bilingues voire multilingues [Degoulet and Fieschi, 1991].

Dans ce contexte, plusieurs ressources terminologiques existent. Citons par exemple l'UMLS [1] (Unified Medical Language System) qui regroupe de nombreuses ressources terminologiques biomédicales permettant l'interopérabilité entre les systèmes informatiques. Elles peuvent être utilisées pour le développement ou l'amélioration de certaines applications sur des textes médicaux telles que la classification documentaire ou les systèmes de traduction.

1.2.2 Pratiques terminologiques en arabe

Les terminologies scientifiques et techniques en arabe sont généralement bilingues ou trilingues, avec la langue anglaise ou française comme langue de référence. Dans la plupart des domaines techniques, nous avons constaté que les termes en arabe standard moderne sont la traduction des termes anglais ou français de ces domaines. Cependant, dans certains domaines, comme la médecine, Wulff [2004] souligne qu'il existe un héritage arabophone important dans la constitution des terminologies.

A l'origine, la création des termes arabes est le résultat de certains évènements historiques. Pendant la période coloniale, les communautés des pays arabes ont été formées principalement en anglais ou en français. Par conséquent, ces dernières sont devenues les langues de communication des spécialistes et des experts, et depuis, la langue arabe n'a pas repris son rôle et sa place comme langue de sciences et de technologies, comme c'était le cas dans l'empire ottoman [Darwish, 2009]. L'impact de cette attitude est difficilement mesurable sur le statut de la langue arabe. Les pays arabes sont devenus désormais importateurs de terminologie [Darwish, 2009].

Dans son histoire, la terminologie arabe a vécu plusieurs changements et a bien évolué. Cette évolution terminologique ne se traduit pas uniquement par la création de termes nouveaux, mais également par la disparition de certains termes souvent longs ou mal adaptés [Hamzé, 2010]. Par exemple, le nom d'instrument **إسم ما عالجت به** (*litt. nom de ce que tu as traité avec*) est remplacé par **إسم الآلة** (*nom d'instrument*). Comme nous l'avons constaté, la plupart des termi-

1. http ://www.nlm.nih.gov/research/umls

nologies actuelles ne présentent pas une création arabe mais plutôt le résultat d'une ouverture sur d'autres terminologies qui proviennent d'autres langues notamment dans le domaine de l'informatique. C'est un domaine où règne l'anglais par excellence. Ceci explique le fait que les termes en MSA dans le domaine de l'informatique dépendent largement des traductions et des translittérations des termes anglais existants. Meyahi [2013] propose de favoriser la création des nouveaux termes arabes que d'emprunter ou adapter des termes étrangers existants car sinon il ne pourra jamais être spontanément compréhensible. De plus, ceci risque de faire reculer de plus en plus la langue arabe.

Dans ce contexte, Hamzé [2010] classifie les pratiques terminologiques arabes en deux catégories. La première est la terminologie grammaticale arabe qui représente une création purement arabe. Celle-ci ne comporte aucun terme étranger comme les terminologies propres à la société arabo-musulmane impliquant les sciences religieuses et les sciences du langage. Manifestement, il est inconcevable d'intégrer des mots étrangers à cette langue pour ces domaines d'étude. La deuxième catégorie inclut les terminologies d'autres sciences, dites étrangères, ayant la traduction à l'origine de leur construction comme la philosophie, la médecine, l'astronomie, etc. Ces termes étrangers, notamment grecs, montrent bien l'influence exercée par la culture d'origine.

Le problème de la normalisation de la terminologie arabe a été véritablement pris en compte lorsque le *Bureau de Coordination de l'Arabisation* (CBA) a été créé à Rabat sous les auspices de l'*Organisation culturelle et scientifique pour l'éducation de la Ligue arabe* (ALESCO) en 1969. Ainsi, le premier vocabulaire scientifique unifié en MSA a été construit en suivant les trois règles suivantes. Premièrement, l'adoption et l'adaptation des mots empruntés ont été découragées étant donné que les termes produits sont parfois inadaptés et que les mots empruntés détruiront « l'esprit » de la langue. Deuxièmement, les experts ont été incités à tirer des termes techniques des racines et des schèmes pour garder le principe de la dérivation pour la formation des nouveaux mots arabes (voir section 1.3.2.2). Troisièmement, en raison de l'arabisation insistante des termes scientifiques, il a fallu déployer beaucoup d'efforts pour trouver des équivalents arabes, souvent longs, pour des termes concis en latin [Emery, 1982].

De nos jours, la situation de la langue arabe dans le domaine de la constitution de terminologie est en progrès même si elle est encore loin d'être idéale. De plus en plus de ressources terminologiques sont développées au sein de projets portant sur la langue arabe. L'objectif est d'automatiser le traitement de la langue arabe, y compris pour le transfert de la terminologie.

Dans le domaine médical, à notre connaissance, la seule ressource terminologique dispo-

nible pour le MSA au format électronique est le MeSH, *Medical Subject Headings*[2]. Il s'agit d'un thésaurus de référence dans le domaine médical. Cependant, la disponibilité des ressources en MSA au format électronique n'est pas toujours garantie. En effet, même si la CIM, la *Classification Internationale des Maladies*[3], est disponible dans les six langues officielles de l'OMS (anglais, arabe, chinois, espagnol, français et russe), seules les versions anglaise et française sont disponibles en format électronique. Ces ressources peuvent être utiles comme données de référence pour évaluer une terminologie acquise automatiquement ou semi-automatiquement. Lors de l'évaluation des différentes méthodes d'acquisition terminologique dans le chapitre 5, nous avons utilisé ce type de ressources comme références.

1.3 L'arabe standard moderne

La langue arabe est une langue afro-asiatique de la famille des langues sémitiques. Elle se divise en deux formes principales : l'arabe dialectal et l'arabe littéraire. De même, il existe deux sortes d'arabe littéraire : l'arabe classique et l'arabe standard moderne (MSA).

1.3.1 Choix de l'arabe standard moderne

L'arabe classique est la langue du Coran et celle utilisée dans les textes anciens. L'arabe standard moderne (اللغة العربية الفصحى, *al-lughat ul-`Arabīyat ul-fuṣḥá*, MSA) est une variante de l'arabe dont la source est l'arabe classique. Elle en présente une évolution littéraire moderne. Tous les pays arabes comprennent le MSA et l'utilisent pour la communication entre eux. En effet, il y représente la langue de la culture depuis des années pour laquelle est développé un énorme répertoire de mots.

Cependant, les dialectes arabes représentent les véritables formes de langue maternelle. Pour communiquer entre eux, les citoyens d'un même pays utilisent leur dialecte. Chaque dialecte arabe est une langue spécifique à son pays. Il varie d'une région à l'autre ou d'un pays à l'autre. Chaque pays, chaque région et chaque ville peut avoir son propre dialecte. Ce dernier représente le moyen de liaison entre les gens pour transmettre les informations, les pensées et les idées. C'est aussi le moyen pour parler aisément et spontanément. Cela facilite leur communication et la rend plus compréhensible. Cependant, comme ils sont principalement parlés non écrits, les dialectes arabes ne sont pas standardisés.

2. http://www.emro.who.int/fr/information-resources/arabic-mesh/
3. http://www.who.int/classifications/icd/en/

Nous avons opté pour le MSA dans notre travail d'acquisition terminologique pour plusieurs raisons. D'abord, il représente la seule variante formelle qui constitue la langue officielle des médias, de la culture et de l'éducation dans le monde arabe alors que les dialectes représentent une variante informelle dont l'utilisation est généralement restreinte pour la communication quotidienne [Habash, 2010]. De plus, ces variétés dialectales sont, de nos jours, extrêmement nombreuses. Surtout, elles ne sont pas utilisées dans les domaines de spécialité car elles ne sont pas suffisamment stables et standardisées.

Étant la langue officielle de 26 pays et de plusieurs organismes internationaux comme l'OMS (Organisation Mondiale de la Santé), de nombreux documents administratifs et techniques sont rédigés en MSA. Il est donc important de disposer de systèmes de gestion terminologique. L'aménagement terminologique est nécessaire dans de nombreux domaines comme l'agriculture, la géologie, la protection de l'environnement ou le droit. Il peut varier d'un pays à l'autre [Massoud, 2003]. De plus, selon l'Internet World Stats 2016, la langue arabe figure en quatrième position parmi les dix premières langues les plus utilisées sur Internet après l'anglais, le chinois et l'espagnol (figure 1.1).

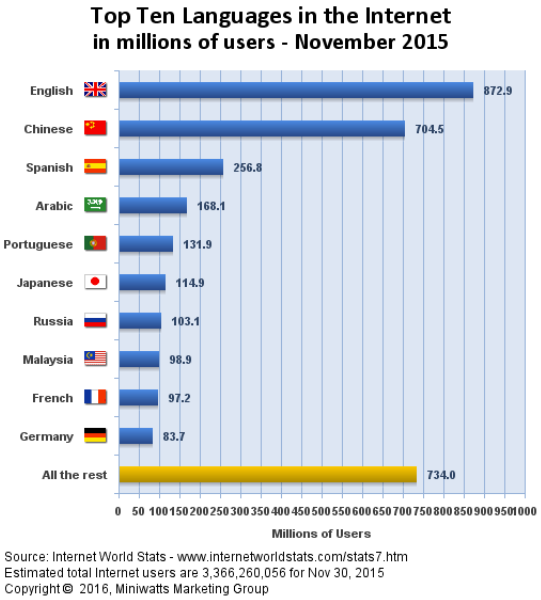


Fig. 1.1 – Les dix langues les plus utilisées sur Internet

Cependant, plusieurs difficultés se manifestent et rendent cette langue difficile à maîtriser dans le domaine du traitement automatique de la langue. Pour cela, nous exposons par la suite quelques phénomènes spécifiques à cette langue.

1.3.2 Spécificités du MSA

L'arabe standard moderne appartient à la famille des langues afro-asiatiques. Plus particulièrement, elle fait partie de la famille des langues sémitiques comme l'hébreu, le phénicien, l'araméen et l'ougaritique. Elle comporte un nombre important de consonnes, 28 en arabe littéral, avec 3 types de diacritiques, voyelles, nunation et shadda, ainsi qu'un système d'écriture de droite à gauche. Cette langue est caractérisée par une morphologie flexionnelle organisée autour de radicaux, principalement verbaux, composés le plus souvent de trois consonnes. Nous exposons ci-dessous certaines caractéristiques linguistiques marquant le MSA comme la voyellation, sa morphologie concaténative des mots, ainsi que sa morphologie dérivationnelle et flexionnelle. Ces particularités rendent difficile l'analyse des données textuelles de notre corpus arabe.

1.3.2.1 Voyellation

L'arabe standard moderne est caractérisé par des signes non alphabétiques qui sont ajoutés au dessus ou au dessous de la lettre. Ces symboles sont appelés des diacritiques.

Selon Habash [2010], il existe trois types de diacritiques que nous positionnons dans les exemples par rapport à une lettre symbolisée par un cercle « ◌ » :

Les voyelles courtes : ce sont des petits sons que l'on ajoute aux consonnes.

Elles sont composées des

i) trois voyelles :

— *Damma* /ou/ ◌ُ ;

— *Fatha* /a/ ◌َ ;

— *Kasra* /i/ ◌ِ ;

Le tableau 1.1 montre la différence entre ces trois voyelles courtes phoniques arabes et les sons qu'ils produisent. En l'absence de l'une des voyelles, c'est la connaissance de la langue qui permet de connaître le sens.

ii) l'absence de toute voyelle marquant l'absence du son : *Soukoun* ◌ْ (prononciation de la lettre sans voyelle).

Cette dernière est représentée par un petit rond au-dessus de la lettre. Elle permet de prononcer le caractère sans son. Les lettres portant un *Soukoun* doivent être précédées par une lettre voyellée. Par exemple, dans بِنْتٌ *binton* (*fille*), la lettre *NOON*[4] **'ن'** *(N)* est prononcée sans voyellation, seul le son de la lettre est prononcé.

4. Les noms des lettres arabes utilisés sont ceux proposés par Unicode.

Voyelle	Nom arabe	Nom en français	Son de la voyelle	Avec consonne	Avec consonne en français
ُ	**ضمّة**	*Damma*	/ou/ comme ‘ou’ dans coup	**بُ**	*bou*
َ	**فتحة**	*Fatha*	/a/ comme ‘a’ dans chat	**بَ**	*ba*
ِ	**كسرة**	*Kasra*	/i/ comme ‘i’ dans lit	**بِ**	*bi*

Tab. 1.1 – Voyelles arabes courtes

La nunation : il s'agit de diacritiques doubles.

Elle apparaît à la position finale des noms, des adjectifs et des adverbes. La présence de ces diacritiques à la fin des noms arabes permet de préciser qu'il s'agit d'un mot indéfini :

— *Dammatan* ٌ (double *Damma*) ;

— *Fathatan* ً (double *Fatha*) ;

— *Kasratan* ٍ (double *Kasra*) ;

Le tableau 1.2 présente la différence entre les trois diacritiques présentant une nunation.

Voyelle	Nom arabe	Nom en français	Son de la voyelle	Mot arabe	Translittération en français
ٌ	**ضمّتان**	*Dammatan*	/on/	**أَلَمٌ**	*alamon*
ً	**فتحتان**	*Fathatan*	/an/	**أَلَمًا**	*alaman*
ٍ	**كسرتان**	*Kasratan*	/in/	**أَلَمٍ**	*alamin*

Tab. 1.2 – Nunation en arabe

La *Shadda* : ّ : c'est un diacritique remplaçant une consonne double.

Il entraîne un doublement de la lettre à laquelle il est accolé. Il est toujours accompagné d’un diacritique simple. Pour la prononciation, il suffit d'insister un peu plus sur la lettre puisqu'en réalité il s'agit de deux lettres identiques dont la première possède un *Soukoun* et la deuxième une voyelle courte autre qu'un *Soukoun*, c'est-à-dire un *Damma*, un *Fatha* ou un *Kasra*. Par exemple, dans le mot **مَرَّ** *MaRRa* (*il est passé*), la lettre *Reh*[5] **'ر'** *(R)* semble être prononcé deux fois, la première avec un *Soukoun* et la deuxième avec une *Fatha*.

L'écriture sans ces diacritiques est dite non voyellée. Cela rend certains mots ambigus : il est alors impossible de décider de leur sens indépendamment du contexte de leur énonciation.

5. Les noms des lettres arabes utilisés sont ceux proposés par Unicode.

Le tableau 1.3 présente les différentes formes et sens que peut avoir le mot **فتح** *F-T-H* en lui apportant différents diacritiques.

Mot arabe	Translittération	Glose
فَتَحَ	*FaTaHa*	***il a ouvert***
فُتِحَ	*FouTiHa*	***il a été ouvert***
فَتْحٌ	*FaTHon*	***ouverture***
فَتَّحَ	*FaTTaHa*	***il a éclos***
فُتَحٌ	*FouTaHon*	***fontes***
فُتُحٌ	*FouTouHon*	***conquêtes***

Tab. 1.3 – Détermination du sens du mot grâce à sa diacritisation

La langue arabe comporte aussi trois voyelles longues. Ce sont les prolongements phonétiques des voyelles brèves, qui conduisent à allonger la durée de leur prononciation. Ces voyelles longues sont :

— le *alif* 'ا' : est toujours précédé d'une lettre portant une *fatha* ◌َ pour prolonger la fatha

— le *waw* 'و' : est toujours précédé d'une lettre portant une *damma* ◌ُ pour prolonger la damma

— le *ya* 'ي' : est toujours précédé d'une lettre portant une *kasra* ◌ِ pour prolonger la kasra

Le tableau 1.4 montre la différence entre les trois différentes voyelles longues arabes.

Voyelle longue	Nom en arabe	Nom en français	Son de la voyelle	Avec consonne en arabe	Avec consonne en français
ا	الف	*alif*	/A/ ou /aa/	بَا	*baa*
و	واو	*waw*	/uw/ ou /ouu/	بُو	*bouu*
ي	ياء	*ya'*	/iy/ ou /ii/	بِي	*bii*

Tab. 1.4 – Voyelles longues arabes

1.3.2.2 Dérivation et flexion

La flexion est la variation de la forme des mots en fonction de facteurs grammaticaux tels que la conjugaison pour les verbes [Hadrich and Chaaben, 2006]. Elle repose sur la concaténation d'affixes à un radical pour construire les différentes formes fléchies. Il existe trois types d'affixes : les préfixes qui précèdent le radical, les suffixes qui succèdent le radical et les circonfixes qui se situent au milieu du radical. Les seuls mots qui restent invariables et ne peuvent pas avoir de formes fléchies sont les particules.

Par exemple, à partir du verbe **يتأثّرون** (*ils s'influencent*), le suffixe **ون** indique le masculin pluriel du verbe **تأثّر** et le préfixe (**ي**) indique le présent pour la troisième personne du singulier masculin, la troisième personne du duel masculin et de la troisième personne du pluriel masculin et féminin.

La dérivation est la formation de nouveaux mots à partir de mots existants. C'est la combinaison d'une racine et d'un schème pour former un radical. Une racine est la suite de trois, quatre ou cinq lettres, dont la plupart sont trilitères, qui définit une notion abstraite. Le schème représente un patron définissant le format du radical qui va suivre. L'obtention des mots dans la langue arabe se fait à partir d'une racine, d'une combinaison de voyelles, de préfixes, d'infixes, de suffixes et d'un *schème* morphologique [Hadrich and Chaaben, 2006]. Par exemple, à partir de la racine **(أ , ث , ر)** composée à partir des trois consonnes *Alef* **'أ'** *(A)*, *Theh* **'ث'** *(TH)* et *Reh* **'ر'** *(R)*, il est possible d'obtenir des verbes comme **تَأَثَّرَ** (*s'émouvoir / être influencé*), des adjectifs comme **متأثّر** (*ému*) et même des noms comme **تَأَثُّرٌ** (*émotion*).

1.3.2.3 Morphologie concaténative d'un mot arabe

La morphologie concaténative en arabe consiste à intégrer des éléments particuliers du lexique, appelés clitiques, au mot auquel ils se rapportent dans un ordre précis [Habash, 2010] :

[QST+ [CNJ+ [PRT+ [DET+ [PRE+ [BASE] +SUF] +PRO]]]]

Ces clitiques représentent des morphèmes optionnels ayant un sens particulier et une fonction syntaxique indépendante. Ils sont à mi-chemin entre un mot et un affixe : ils possèdent les caractéristiques syntaxiques d'un mot, et bien que, comme les affixes, ils soient morphologiquement liés à un autre mot, ils restent grammaticalement indépendants. Ces morphèmes sont généralement des pronoms, des articles, des conjonctions de coordination ou des prépositions. Plusieurs clitiques peuvent être rattachés au même mot arabe suivant l'ordre indiqué précédemment. Ainsi, on distingue deux types de clitiques selon leur position par rapport au mot associé : les proclitiques et les enclitiques.

Les proclitiques sont des éléments lexicaux qui précèdent le mot. Ils peuvent représenter un article (**ال**) (le, la, les), une préposition (**ل, ب**) (pour, par), une conjonction (**و**) (et), etc. Le tableau 1.5 résume les proclitiques arabes les plus utilisés [Habash, 2010].

Les enclitiques arabes sont des éléments qui suivent le mot. Ils expriment uniquement des pronoms. Ils représentent un pronom possessif si le mot auquel ils sont rattachés est un nom et ils désignent le complément d'objet s'il s'agit d'un verbe ou d'une particule. Contrairement aux proclitiques, les enclitiques varient en genre et en nombre. Plusieurs clitiques peuvent apparaître

Clitique	Classe	Fonction	Glose
أ *Â*	QST	particule interrogative	est-ce que
و *wa*	CNJ	Conjonction de coordination (entre deux groupes nominals)	et
و *wa*	CNJ	Conjonction de connexion (entre deux phrases)	et
و *wa*	CNJ	Conjonction de coordination	et
و *wa*	CNJ	Conjonction de subordination de circonstancié	en
و *wa*	PRT	préposition de serment	par
و *wa*	PRT	préposition d'accompagnement	avec
ف *fa*	CNJ	conjonction	et / et alors
ف *fa*	CNJ	Particule de connexion	et / et alors
ف *fa*	CNJ	Particule de réponse conditionnelle	et / et alors
ف *fa*	CNJ	Conjonction de subordination conditionnelle	alors
ب *bi*	PRT	Préposition	par / en
ك *ka*	PRT	Préposition	comme
ل *li*	PRT	Préposition	à, pour, en faveur de, à cause de
س *sa*	PRT	Particule du futur	aller + [infinitif]
ال *al*	DET	Article défini	le

Tab. 1.5 – Liste des proclitiques arabes les plus utilisés

dans un seul mot arabe. Ce mot peut alors correspondre à toute une expression en français. Par exemple, le mot **وسيكتبونها** *wasayaktubuwnahA* se traduit par l'expression française *et ils vont l'écrire*. La segmentation d'un tel mot (exemple 1) s'avère à la fois nécessaire et difficile pour l'analyse sémantique : "**و**" *wa* - *proclitique* (*et*) + "**س**" *sa* - *proclitique* (*pour indiquer le futur*) + "**ي**" *ya* (*pour indiquer que le sujet est à la troisième personne*) + "**كتب**" *ktub* (*le verbe écrire / écrit*) + "**ون**" *uwna* (*pour indiquer que le sujet est au masculin pluriel*) + "**ها**" *hA* - *enclitique* (*pronom personnel à la troisième personne féminin singulier*).

(1) وسيكتبونها

و س ي كتب ون ها

wa sa ya ktub uwna hA

et FUT SUJ3 écrire SUJmp PRO3fs

et ils vont l'écrire

Dans cet exemple, la base du mot est **يكتبون** - *yaktubuwna* (*ils écrivent*). Trois clitiques lui

sont associés ; i) deux proclitiques à savoir la conjonction de coordination و - *wa* et la particule س - *sa* exprimant le future ainsi que ii) le pronom personnel de la troisième personne féminin singulier ها - *hA* comme un enclitique.

1.3.3 Traitement automatique de la langue arabe

Lors de l'analyse automatique de textes rédigés en arabe, plusieurs difficultés se manifestent et rendent cette langue difficile à traiter automatiquement. Comme nous l'avons déjà signalé dans la section précédente, certaines sont liées aux caractéristiques intrinsèques de la langue elle-même comme sa morphologie concaténative des mots (1.3.2.3) ou l'absence de voyellation marqué par la quasi-totalité des textes arabes (1.3.2.1). De plus, il s'agit d'une langue qui s'écrit de droite à gauche ce qui perturbe le fonctionnement de certains outils, ce que nous détaillons par la suite à la section 3.2.3.

Même si cette langue se caractérise par la présence de diacritiques, la plupart des textes arabes ne sont pas voyellés. Dans certains cas, ceci engendre des ambiguïtés lors de l'analyse des textes. Par exemple, cela peut entraîner des erreurs lors de l'étiquetage morpho-syntaxique des textes. Dans ألم متوسط (*douleur modérée*), le mot ألم *Alam* désignant le mot ***douleur*** est analysé par l'outil MADA+TOKAN ainsi que MADAMIRA (voir section 3.2.4) comme suit :

(2) ألم

أ	لم
A	lam
particule-interrogative	négation-au-passé

est-ce que ne pas

Cette ambiguïté aurait dû être résolue si le texte était diacritisé. S'il s'agissait d'un pronom interrogatif, le mot ألم aurait obligatoirement un *Soukoun* ◌ْ (prononciation de la lettre sans voyelle) au dessus de sa dernière lettre (en prononçant *alam*) car la particule de négation لَمْ *lam* (pour exprimer la négation d'un verbe à sens passé) ne peut figurer qu'avec cette voyellation. De plus, il devrait précéder un verbe au présent. Dans notre cas, comme il s'agit d'un nom, il ne peut jamais avoir un *Soukoun* à la fin, le dernier caractère des noms arabes ne portant jamais cette diacritique. Si le texte était diacritisé, ce mot aurait porter une nunation أَلَمٌ *Alamon*.

La morphologie concaténative des mots étant une caractéristique élémentaire de la langue arabe, appelée aussi *agglutination*, peut, pour sa part, causer des erreurs au cours de l'étiquetage

morpho-syntaxique des textes. Du point de vue du traitement automatique de l'arabe, il est parfois difficile de distinguer un proclitique ou enclitique du reste du mot en question. Revenons à l'exemple précédent, comme il est possible d'associer différents éléments à un mot, les outils MADA+TOKAN et MADAMIRA (voir section 3.2.4) ont confondu entre le pronom interrogatif أ -- *A* (*est-ce que*) et le premier caractère du mot ألم (*douleur*). L'ambiguïté est plus importante lorsque les diacritiques ne sont pas représentés. Nous allons rencontrer ces problèmes lors de la mise au point des différentes méthodes d'extraction terminologique à partir des textes arabes (voir chapitre 4).

D'autres freins au traitement automatique de l'arabe sont dûs au manque et à la déficience de ressources et d'outils adaptés à cette langue. Dans le domaine médical et de la santé, l'arabe présente certaines difficultés relatives aux ressources et aux outils linguistiques. Ces difficultés sont dues à certaines pratiques adoptées dans de nombreux pays à travers le monde arabe (voir section 1.2.2).

De plus, les langues ne sont pas outillées de la même manière. Certaines langues telles que l'arabe standard moderne manquent de ressources linguistiques et terminologiques. Les difficultés de développement de telles méthodes du TAL sont également inhérentes aux caractéristiques intrinsèques de la langue elle-même. Ainsi ces barrières freinent la mise en œuvre de méthodes d'analyse linguistique et terminologique. Cela explique le fait que plusieurs méthodes ont plus souvent été proposées pour l'extraction de termes pour la langue anglaise ou d'autre langues européennes.

1.4 Problématique

Pour diverses raisons, la plupart des approches et des outils de TAL sont essentiellement destinés aux principales langues d'Europe et surtout à l'anglais. C'est également le cas dans le domaine de la terminologie. Cependant, de nos jours, ce constat tend à évoluer avec des conséquences sur les pratiques mises en œuvre en terminologie computationnelle.

La majorité des outils de traitement automatique des langues développés pour l'acquisition terminologique se focalisent sur ces langues marquées comme « bien traitées ». Mais, toutes les langues ne sont pas outillées de la même manière. Les langues telles que l'arabe standard moderne souffrent d'une carence de ressources et d'outils disponibles dans le domaine d'acquisition terminologique.

Aussi, comme nous l'avons marqué à la section 1.3.2, les difficultés de développement de telles ressources de TAL sont également inhérentes aux caractéristiques intrinsèques de la

langue elle-même comme la voyellation, la morphologie dérivationnelle et flexionnelle de la langue ainsi que la morphologie concaténative des mots arabes. Ces barrières freinent ainsi la mise en œuvre d'extracteurs terminologiques pour l'arabe.

1.4.1 Objectifs

L'objectif de notre travail est de lever les verrous que constituent le manque de disponibilité de ressources ou d'outils TAL pour la langue arabe dans les domaines de spécialité en proposant des méthodes permettant l'extraction de termes à partir de textes en arabe standard moderne.

Ainsi, dans notre thèse, nous souhaitons répondre à la question de recherche suivante : Si une langue ne dispose de méthode ou d'outil pour une tâche donnée, dans notre cas l'extraction terminologique, est-ce qu'il vaut mieux i) adapter un outil existant, ii) mettre en oeuvre une méthode de transfert d'une langue à l'autre, donc translingue, ou iii) mettre au point une méthode spécifique monolingue.

1.4.2 Contributions

Nos contributions portent sur l'extraction terminologique pour l'arabe standard moderne (MSA) dans le domaine médical à partir des corpus monolingues arabes et bilingues parallèles. Notre première contribution consiste en la construction d'un corpus parallèle dans un domaine de spécialité. Il s'agit d'un ensemble de textes médicaux produits par la *United States National Library of Medicine* (NLM) présentant des brochures à destination des patients. La préparation de ce corpus parallèle aligné anglais-arabe a nécessité de mettre en œuvre des traitements répartis en plusieurs étapes.

Pour l'extraction terminologique en MSA, nous proposons trois axes de travail. Dans un premier temps, nous proposons une méthode d'extraction des termes en effectuant une adaptation de l'extracteur terminologique YATEA [Aubin and Hamon, 2006] à l'analyse de textes en MSA. Cet extracteur est initialement mis au point pour extraire des termes à partir de textes de spécialité en français ou en anglais. Nous présentons la méthodologie d'adaptation permettant de prendre en compte, dans le processus d'extraction de termes, certaines spécificités morphologiques et morpho-syntaxiques décrites à la section 1.3.2.

Dans un deuxième temps, nous nous sommes intéressés à l'assimilation de mots spécialisés pour la langue arabe. Nous proposons ainsi une méthode qui permet d'extraire des termes arabes à partir de la translittération des termes anglais en caractères arabes. Pour cela, nous avons créé une table de correspondance des caractères anglais vers l'arabe. Cette méthode proposée prend

en compte certaines particularités de la langue arabe comme l'agglutination et la non voyellation des textes arabes.

Enfin, nous proposons une méthode d'acquisition terminologique basée sur la notion de transfert translingue [McDonald et al., 2011] appliquée à l'arabe. Pour ce faire, nous nous appuyons sur un corpus parallèle anglais-arabe dans le domaine médical et un transfert des termes extraits de textes en anglais vers la langue arabe sans utiliser un extracteur pour l'arabe. Une étape d'alignement et une extraction terminologique à partir du corpus anglais sont indispensables à la mise en place de notre méthode de transfert.

1.4.3 Schéma récapitulatif

La figure 1.2 présente sous forme schématique les méthodes proposées dans la thèse ainsi que les corpus utilisés et les résultats produits.

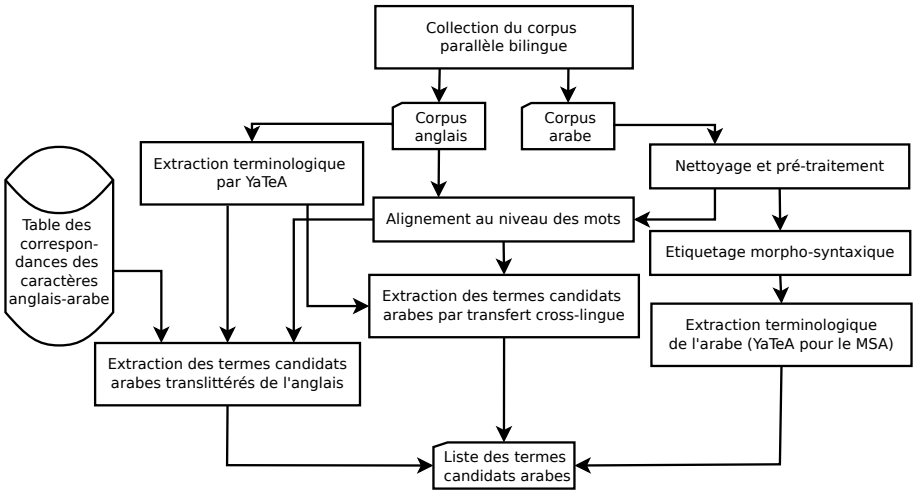


Fig. 1.2 – Schéma récapitulatif des méthodes proposées dans la thèse.

1.5 Organisation du document

Ce manuscrit est organisé en six chapitres. Après ce premier chapitre introductif, nous présentons dans le chapitre 2 un inventaire sur les différentes méthodes existantes pour l'extraction terminologique. Nous décrivons dans un premier temps les principes généraux de l'extraction terminologique ainsi que quelques outils monolingues les mettant en œuvre. Par la suite, nous présentons les méthodes d'acquisition terminologique sur les textes en arabe standard moderne (MSA) ainsi que celle sur les textes multilingues. Enfin, nous proposons une revue de travaux de translittération qui se sont focalisés sur la langue arabe.

Le chapitre 3 est consacré à la présentation de notre corpus de travail. Tout d'abord, nous décrivons les différents traitements mis en œuvre dans la collecte des documents et la constitution de notre corpus parallèle anglais-arabe du domaine médical. Ensuite, nous détaillons l'étape d'alignement effectué en utilisant outil GIZA++.

Dans le chapitre 4, nous proposons plusieurs stratégies pour l'extraction terminologique à partir de textes arabes. Nous exposons l'adaptation de l'extracteur terminologique YATEA pour l'arabe. Puis, nous proposons une méthode qui permet de repérer et d'extraire des couples de termes médicaux produits à partir de la translittération des termes anglais en caractères arabes. Finalement, nous définissons une méthode d'acquisition terminologique pour le MSA utilisant une méthodologie de transfert translingue.

Nous abordons dans le chapitre 5 l'évaluation des trois méthodes proposées pour extraire des termes à partir de textes arabes. Nous décrivons le protocole d'évaluation mis en place pour procéder à la validation des termes monolingues et des couples de termes candidats anglais-arabe. Mais avant cela, nous présentons l'évaluation des alignements effectués pour notre corpus parallèle anglais-arabe. Puis, nous décrivons et analysons les résultats obtenus par chacune des méthodes proposées et nous dressons un bilan afin de répondre à la question de recherche posée dans la section 1.4.1.

Finalement, au chapitre 6, nous concluons en récapitulant les différentes contributions et en proposant des perspectives à notre travail.

Chapitre 2

Etat de l'art

Sommaire

2.1 Introduction

Les méthodes de Traitement Automatique du Langage Naturel (TAL) sont incontournables pour identifier et extraire les termes dans les textes de spécialité. Les informations issues de l'analyse des textes facilitent la mise en œuvre d'applications destinées à des utilisateurs finaux, comme par exemple la recherche d'information, la traduction automatique ou l'extraction d'informations. Dans les domaines de spécialité, les méthodes de TAL s'appuient sur des ressources terminologiques constituées généralement à partir de corpus textuels. Les méthodes d'extraction automatique de termes facilitent la création de ressources terminologiques adaptées aux textes traités. Aussi, étant donné la masse importante de données à analyser, la tâche de constitution de terminologie ne peut être envisagée qu'avec le recours à des outils informatiques d'acquisition terminologique [Boulaknadel et al., 2008 ; Aubin and Hamon, 2006].

Dans ce chapitre, nous nous intéressons aux différents axes abordés dans notre travail de thèse. Dans la section 2.2.1, nous présentons les principes généraux de l'extraction terminologique ainsi que les premiers outils d'acquisition terminologique. Dans la section 2.2.2, nous nous intéressons à l'acquisition terminologique sur les textes en arabe standard moderne (MSA), puis celle sur les textes multilingues dans la section 2.2.3. Finalement, nous exposons certains travaux de translittération dans la section 2.3.

2.2 Acquisition terminologique

Depuis les années 90, de nombreux travaux de recherche ont conduit à la mise au point de méthodes d'acquisition terminologique à partir de textes de spécialité (articles scientifiques, documentation technique, textes juridiques, etc.) afin d'assister le travail de constitution de ressources terminologiques par les terminologues [Cabré et al., 2001 ; Pazienza et al., 2005]. Ces résultats permettent également de prendre en compte la terminologie d'un domaine dans les applications facilitant ainsi l'accès à l'information spécialisée contenue dans les textes [Marshman et al., 2012 ; Cohen and Demner-Fushman, 2013]. En effet, l'acquisition terminologique est une tâche indispensable pour l'accès aux informations présentes dans les textes de spécialité dans le but de construire une terminologie d'un domaine spécifique. Il s'agit d'un processus semi-automatique. Tout d'abord, les termes candidats sont identifiés et extraits automatiquement. Ces listes de termes doivent être être ensuite validés par les terminologues. Étant donné un domaine d'activité, il existe plusieurs terminologies qui peuvent représenter les connaissances de ce domaine. En effet, les terminologies d'un domaine doivent prendre en compte l'application

et l'utilisateur final. Du point de vue des termes retenus et de leur description, celles-ci peuvent varier selon le besoin.

2.2.1 Principes généraux

À la croisée de la terminologie traditionnelle et du traitement automatique des langues (TAL) se situe la terminologie computationnelle. Celle-ci vise à produire des ressources terminologiques à partir de corpus en s'appuyant sur des outils d'analyse automatique des textes. Les premiers outils ont été développés essentiellement pour la langue française et anglaise dans un contexte industriel. Le premier outil d'acquisition terminologique à partir de corpus, appelé TERMINO [David and Plante, 1990] (devenu aujourd'hui NOMINO[1]), a été développé à l'Université du Québec à Montréal. Cet outil, dédié à la construction de bases terminologiques, a été développé dans le cadre d'un projet soutenu par l'Office de la Langue Française (OLF) du Québec. Par la suite, d'autres travaux ont vu le jour comme LEXTER [Bourigault, 1994], un logiciel d'extraction terminologique, ANA [Enguehard and Pantera, 1995] développé pour l'enrichissement de réseaux lexicaux exploités par un système de gestion des connaissances, ACABIT [Daille, 1994] pour la construction de lexiques terminologiques multilingues,... Certains d'entre eux se basent sur un corpus de textes préalablement étiquetés morpho-syntaxiquement [Bourigault, 1994 ; David and Plante, 1990 ; Daille, 1994].

Bien que les méthodes distributionnelles ou de classification supervisées peuvent être utilisées [Nazar et al., 2012 ; Zadeh and Handschuh, 2014 ; Conrado et al., 2013], les approches s'appuient principalement sur une description linguistique du processus d'extraction et des filtres statistiques [Bourigault, 1993 ; Daille, 2003 ; Drouin, 2002 ; Aubin and Hamon, 2006]. En effet, l'extraction terminologique repose sur deux étapes : elle consiste d'abord à extraire automatiquement un ensemble de termes candidats à partir du corpus qui sont par la suite fournis à une étape de prise de décision afin de retenir les termes. Afin de mesurer la qualité des termes candidats extraits et décider s'il s'agit bien d'un terme du domaine, deux indices sont attribués à chaque terme [Kageura and Umino, 1996] :

— le figement (*Unithood*) : Il s'agit du degré de stabilité des combinaisons syntagmatiques ou des collocations ;

— le potentiel terminologique (*Termhood*) : Cette caractéristiques décrit la force de la relation que les termes entretiennent avec un domaine de spécialité.

1. http ://www.ling.uqam.ca/nomino/

La notion de *unithood* est généralement associée à celle de *termhood*. Ces deux notions constituent la base de nombreuses méthodes d'extraction terminologique [Zhang et al., 2008]. L'extraction terminologique nécessite d'abord de repérer les séquences stables dans le corpus grâce à des méthodes linguistiques ou symboliques. La qualité est mesurée par le premier indice (*unithood*). Puis, le deuxième indice (*termhood*) permet de mesurer de la distance du terme avec le domaine et l'application en se basant sur méthodes statistiques.

Plusieurs stratégies d'identification et d'extraction de terme ont été proposées. Elles utilisent principalement des approches linguistiques identifiant des syntagmes nominaux pouvant être ou contenir des termes. Il s'agit de définir, identifier et reconnaître les termes en s'appuyant sur leurs propriétés linguistiques comme les traits morpho-syntaxiques et les lemmes [Benveniste, 1966 ; Aubin and Hamon, 2006 ; Korenchuk, 2014] ou d'identifier des modèles syntaxiques de termes en utilisant des techniques de filtrage linguistique comme les patrons syntaxiques [Earl, 1970 ; Aubin and Hamon, 2006 ; Marrero García et al., 2015]. Bourigault [1994] utilise une méthode linguistique par repérage de frontières syntaxiques (verbes conjugués, pronoms, conjonctions de subordination, etc.) pour l'extraction de syntagmes nominaux maximaux dans la mise en œuvre de l'outil LEXTER. Par la suite, des règles syntaxiques sont appliquées sur les syntagmes nominaux maximaux extraits et produisent les termes candidats proposés par le logiciel. Dans le but d'extraire des termes candidats binaires, Daille [1994] identifie ses termes grâce à l'utilisation d'un ensemble de séquences d'étiquettes mises en œuvre sous forme de transducteurs : `Nom Adj`, `Nom1 à (Det) Nom2`, etc. D'autres travaux se sont basés sur les marqueurs lexico-syntaxiques [Daille, 2003] ou utilisent une analyse syntaxique de surface basée sur les frontières syntaxiques et l'analyse endogène [Bourigault, 1993]. TERMINO effectue une analyse morpho-syntaxique afin de repérer les termes candidats en se basant sur des règles morphologiques ainsi qu'une analyse des collocations nominales à l'aide d'une grammaire. En revanche, Enguehard and Pantera [1995] traitent des textes écrits et des interviews transcrits pour l'extraction des termes sans faire recours à une analyse linguistique. Leur système exploite une liste de termes de référence pour analyser leur structure afin de rechercher des structures similaires.

Dans une deuxième étape de classement des termes candidat s'appuyant principalement sur des filtres statistiques, plusieurs mesures ont été utilisées pour l'étape de validation des termes. Certains travaux s'appuient sur une seule mesure statistique. Bourigault [1994] utilise le calcul du *coefficient de productivité* pour mesurer la densité du réseau autour du candidat terme. D'autres combinent une ou plusieurs mesures. Enguehard and Pantera [1995] utilisent

différentes mesures parmi lesquelles la mesure *Log-Likehood Ratio (LLR)* (rapport de vraisemblance) Dunning [1993] afin de retenir les meilleurs termes candidats sans être sensible au nombre d'occurrences ainsi que le calcul de l'*Information Mutuelle (IM)* comme mesure d'association lexicale. Testé sur un corpus anglais composé d'environ 25 000 mots avec 29 termes de référence, cette méthode a permis d'extraire 200 termes mais ayant avec un taux d'erreur de 25%.

La plus importante des mesures utilisées pour la détection des termes est *la fréquence* (nombre d'occurrences des termes) [Daille, 1994] bien que celle-ci influe sur les résultats en identifiant un nombre important de séquences fréquentes qui ne représentent pas des termes. En revanche, certains termes qui apparaissent rarement ne sont pas extraits. Pour cela, ce système s'appuie aussi sur d'autres mesures comme le critère de vraisemblance. De son côté, grâce à son mécanisme d'apprentissage endogène qui assure son fonctionnement de manière autonome, TERMINO [David and Plante, 1990] arrive identifier entre 70% et 74% des termes complexes. Ses erreurs sont dues à la coordination (marquant la rupture de segment textuel), la présence des acronymes et des noms en majuscules.

Dans la plupart des publications et des ouvrages traitant d'extraction terminologique, les termes définis sont de nature nominale ([Sager, 1990 ; Daille, 1994 ; Justeson and Katz, 1995]). Comme nous l'avons déjà signalé, un terme peut être simple, constitué d'un seul mot ou d'une seule unité lexicale, ou complexe, constitué de plusieurs unités lexicales ou d'une unité phraséologique. Cependant, en raison de leur structure syntaxique réduite à une seule unité lexicale, les termes simples sont plus ambigus hors de leurs contextes et plus difficile à extraire automatiquement. Par contre, les termes complexes posent moins de problèmes de polysémie et ils peuvent être extraits à l'aide d'une analyse syntaxique superficielle. Celle-ci permet d'obtenir les termes qui sont eux-mêmes composant de termes plus complexes. La plupart des extracteurs de termes se focalisent sur les termes complexes car ils sont souvent très majoritaires dans les terminologies. Dans un premier temps, les séquences extraites du corpus suite à une analyse sont appelées des termes candidats, c'est-à-dire des syntagmes nominaux susceptibles d'être retenus comme des termes. Ils n'acquièrent le statut de termes qu'après validation par un expert ou un terminologue Bourigault et al. [2004]. Cette décision est prise en fonction des objectifs de l'application. Ces termes candidats doivent présenter des caractéristiques propres au domaine et correspondre aux objectifs de l'application et au corpus sur lequel elle repose.

2.2.2 Extraction terminologique en arabe

L'intérêt croissant pour le traitement automatique de la langue arabe a conduit à mettre en oeuvre des méthodes d'extraction de termes sur cette langue. Celles-ci utilisent des approches similaires à celles réalisées sur l'anglais ou le français Bourigault [1993] ; Daille [2003] ; Drouin [2002] ; Cabré et al. [2001] ; Pazienza et al. [2005] ; Aubin and Hamon [2006].

Cependant, les méthodes d'extraction de termes à partir de textes en arabe sont plus rares. Ce constat peut s'expliquer par la complexité de cette langue : les approches traditionnelles d'acquisition terminologique ne prennent pas en compte plusieurs phénomènes linguistiques comme l'absence de voyellation, l'agglutination et les ambiguïtés morphologiques et syntaxiques des phrases nominales Boulaknadel et al. [2008]. Une autre explication réside dans le fait que la lange arabe est un ensemble de variantes linguistiques incluant l'arabe standard moderne (MSA), l'arabe classique, et de nombreux dialectes Habash [2010]. Aussi, comme nous l'avons signalé, dans de nombreux domaines de spécialité comme la médecine, le français ou l'anglais est la langue utilisée lors de la pratique et de l'enseignement Samy et al. [2012].

A l'instar des méthodes mises en œuvre pour extraire des termes sur le français ou l'anglais, l'extraction de termes en arabe combine une description linguistique du processus d'extraction et des filtres statistiques pour ordonner les termes extraits tout en tenant compte des spécificités de l'arabe. Bounhas and Slimani [2009] proposent d'extraire des termes complexes candidats à l'aide d'une approche hybride composée de deux étapes. Un premier filtre linguistique exploite les résultats de l'analyse morphologique et l'étiquetage morpho-syntaxique des textes pour identifier des séquences de mots candidates. Un second filtre statistique utilisant comme mesure d'association, le rapport de vraisemblance *Log-Likelihood Ratio* (LLR) Dunning [1993] est appliqué sur les résultats ambigus de la première étape pour sélection la meilleure solution.

AlKhatib and Badarneh [2010] proposent une approche hybride similaire à la précédente mais utilisent deux mesures statistiques : i) le LLR pour identifier le degré de stabilité de la combinaison syntagmatique candidate (*unithood*) ; ii) la C-Value [Frantzi and Ananiadou, 1997] pour calculer le degré de liaison de l'unité terminologique au domaine spécifique (*termthood*). Les deux approches ont été évaluées sur un corpus de textes en arabe du domaine de l'environnement, issus de sites Web. Les résultats ont montré que l'utilisation de la méthode C-Value donne de meilleurs résultats que la méthode LLR. Ainsi, l'utilisation de la combinaison des deux méthodes leur donne les meilleurs résultats

Abed et al. [2013] ont adapté des méthodes destinées à l'analyse de textes de la langue générale pour analyser des textes d'un domaine spécifique (des textes religieux) et ainsi ex-

traire automatiquement les termes simples et complexes du domaine. Un corpus électronique contenant de l'arabe classique et de l'arabe standard moderne, collecté à partir des archives de journaux islamiques et des sites islamiques, est utilisé pour évaluer l'approche. Dans ce travail, plusieurs mesures (information mutuelle, Kappa, χ^2, T-test, Piatersky-Shapiro et l'agrégation des rang) sont utilisés pour calculer le degré d'association des composants des termes complexes (*unithood*), et le TF*IDF est utilisé pour ordonner les termes simples en fonction de leur *termhood*.

Comme le souligne Bounhas et al. [2014], l'évaluation des approches proposées est assez critiquable : seuls quelques centaines de termes classés parmi les premiers, sont évalués manuellement, alors que plusieurs milliers ont pu être extraits et que, quelle que soit l'approche et les mesures utilisées, les résultats sont généralement de bonne qualité lorsqu'on ne tient compte que des premiers termes [Korkontzelos et al., 2008 ; Hamon et al., 2014]. Ceci peut s'expliquer par la difficulté intrinsèque à disposer de références pour l'évaluation des méthodes d'acquisition terminologique (un constat similaire peut être fait dans bien d'autres langues et notamment le français) combinée aux manques de disponibilité de collection de textes en arabe issus de domaine de spécialité. Notons également que les systèmes présentés ci-dessus ne sont pas librement accessibles : il n'est donc pas possible de reproduire ou de comparer les résultats.

2.2.3 Extraction terminologique multilingue

Des travaux se sont intéressés à l'acquisition terminologique à partir de textes spécialisés rédigés en arabe moderne standard, principalement pour acquérir des termes complexes 2.2.2, alors que d'autres se sont intéressés à l'acquisition terminologique bilingue ou multilingue à partir de corpus parallèles [Fan et al., 2009 ; Fawi and Delmonte, 2015 ; Kontonatsios et al., 2014] ou comparable [Daille, 2012 ; Fung and Mckeown, 1997] composés de textes spécialisés.

Pour la plupart des travaux qui utilisent des corpus parallèles, la construction de terminologie multilingue se base sur deux étapes. D'abord, les termes candidats sont extraits pour chaque langue. Par la suite, des correspondances sont établies entre les termes des langues différentes en utilisant soit un outil d'alignement au niveau des mots [Fan et al., 2009], soit un outil d'alignement au niveau des séquences [Ideue et al., 2011] comme pour la traduction automatique.

Dans cette section, nous nous intéressons aux travaux s'appuyant sur des corpus parallèles. Ainsi, l'extraction monolingue présente une étape intermédiaire pour la construction de liste de termes multilingues. Dans ce contexte, Fawi and Delmonte [2015] proposent une approche d'extraction des termes complexes à partir d'un corpus parallèle italien-arabe de textes juridiques.

Il s'agit d'une approche hybride qui combine des connaissances linguistiques et statistiques. Tout d'abord, une analyse morpho-syntaxique est effectuée pour les deux corpus en utilisant les étiqueteurs AMIRA [Diab et al., 2007] pour les textes arabes et VEST [Delmonte, 2007] pour les textes italiens. L'extraction des termes candidats monolingues se base alors d'une part sur les étiquettes morpho-syntaxiques associées aux mots du corpus, et d'autre part sur les patrons syntaxiques proposés par Mahdaouy et al. [2013] pour l'arabe et ceux proposés par Bonin et al. [2010] pour l'italien. Des filtres statistiques sont utilisés comme le *Log-Likelihood Ratio* (LLR) pour ordonner les termes candidats en mesurant le degré de stabilité de la combinaison syntagmatique candidate *(unithood)*, ainsi que la méthode C-NC value [Frantzi et al., 2000] qui combine des informations linguistiques et statistiques pour calculer le degré de liaison de l'unité terminologique par rapport au domaine *(termhood)*. La deuxième étape consiste à extraire et établir des liens de correspondances entre les termes monolingues extraits de ces corpus multilingues. Dans le travail de Fawi and Delmonte [2015], les termes anglais et arabes sont annotés dans le corpus. En utilisant un alignement au niveau des phrases, les traductions équivalentes des termes candidats sont identifiés. Après avoir regroupé chaque unité de traduction dans un dictionnaire des termes bilingues, le système valide les équivalents de traduction réels. La validation de ces traductions est basée sur trois mesures : i) la valeur du LLR pour estimer le degré d'association entre les termes complexes bilingues, ii) l'utilisation d'un système de traduction automatique statistique notamment Google Translate et iii) l'utilisation d'index des termes complexes dans le contexte parallèle comme indicateur de relation de traduction.

Afin de construire une terminologie bilingue arabe-anglais de termes simples [Lahbib et al., 2014] s'appuient sur un corpus parallèle de textes religieux dont la plupart sont diacritisés. L'utilisation de l'outil de désambiguïsation probabiliste Ayed et al. [2012] pour les textes arabes permet la segmentation des phrases en mots, la suppression des éléments flexionnels ainsi que l'attribution de la partie du discours à chaque mot. Les textes anglais sont étiquetés en utilisant l'outil TreeTagger [Schmid, 1997]. L'extraction des termes candidats simples est effectué uniquement sur les textes arabes. En plus de l'étiquette morphologique, des mesures statistiques sont utilisées comme la mesure TF-IDF [Jones, 1972] ou la mesure de Lafon [Lafon, 1980]. Un alignement au niveau des mots est effectué à partir des textes parallèles arabe-anglais en utilisant l'outil GIZA++ [Och and Ney, 2003]. La dernière étape consiste à construire une matrice de traduction à partir du processus d'alignement pour identifier les alignements corrects en utilisant les co-occurrences de chaque paire de mots des deux langues ainsi que leur étiquette

morphologique. Pour évaluer l'approche, un corpus composé d'un ensemble de hadith[2] voyellés est utilisé. L'application de la méthode sur ces données permet d'obtenir un taux de réussite de 90%. Les bons résultats obtenus par cette méthodes sont dus à deux facteurs. D'une part, l'approche proposée ainsi que l'évaluation sont réalisées sur des corpus arabes diacritisés ce qui est reconnu pour faciliter l'analyse des textes en langue arabe. D'autre part, la plupart des couples de termes extraits présentent des translittérations des termes arabes pour l'anglais.

D'autres travaux comme celui de Hamon and Grabar [2016] se sont focalisés sur des langues peu dotées comme l'ukrainien. L'objectif est de construire une terminologie bilingue et trilingue dans le domaine médical. Cette méthode se base sur un corpus multilingue collecté à partir de MedlinePlus. Il s'agit d'un ensemble de textes parallèles et alignés ukrainien/anglais/français. L'extraction terminologique monolingue est effectuée sur les textes anglais et français en utilisant l'outil YaTeA [Aubin and Hamon, 2006]. Une méthode de transfert de la langue source (l'anglais ou le français) vers la langue cible (l'ukrainien) est ensuite mis en œuvre pour construire des liens de correspondance entre les termes monolingues extraits (en anglais et en français) et leurs traductions dans les textes ukrainien. Cette méthode utilise à la fois la structure du document et les termes candidats extraits des textes de la langue source qui sont transférés par la suite vers la langue cible en se basant sur l'alignement fourni par GIZA++.

De même, afin de construire et de mettre à jour des dictionnaires bilingues des termes médicaux pour des langues peu outillées comme le grec et le roumain, Kontonatsios et al. [2014] proposent une approche hybride visant à extraire des termes complexes à partir du corpus biomédical multilingue EMEA parallèle anglais-grec-roumain. L'extraction des termes de la langue source consiste à analyser le corpus anglais pour identifier les termes de l'UMLS. L'identifiant du concept UMLS (CUI - *Concept Unique Identifier*) et la catégorie sémantique (également issue de l'UMLS) sont associés à chacun des termes. Pour obtenir l'équivalence de traduction dans la langue cible, Kontonatsios et al. [2014] réalisent différents d'alignement de termes anglais extraits : i) un module d'alignement d'expressions utilisant GIZA++ pour classer des paires d'expressions candidates ; ii) une utilisation du *Random Forest (RF) aligner*, cet aligneur étant basé sur une approche d'apprentissage supervisé utilisant des critères n-grammes pour la génération d'une liste des **N* meilleures traductions candidates ; iii) un système de vote basé sur l'intersection des deux méthodes précédentes, le but étant d'augmenter la précision des dictionnaires extraits automatiquement. L'identifiant du concept et la catégorie sémantique sont

2. C'est une communication orale du prophète de l'islam Mahamed et, par extension, un recueil qui comprend l'ensemble des traditions relatives aux actes et aux paroles de Mahamed et de ses compagnons, précédées chacune d'une chaîne de transmetteurs remontant jusqu'à Mahomet.

propagés du terme source à la traduction cible correspondante. Une évaluation manuelle est effectuée sur 1000 termes anglais tirés aléatoirement. Les 20 meilleures traductions candidates sont sélectionnées pour chacun des termes. Le module d'alignement d'expressions produit de meilleures performances par rapport à celles obtenues avec le *RF aligner*, pour les couples de termes anglais-grecques ainsi que pour les couples de termes roumains-grecques.

2.3 Translittération des mots

Plusieurs travaux de recherche ont été menés ces dernières années sur la translittération concernant l'arabe. Les travaux se sont principalement focalisés sur le passage de l'arabe vers l'écriture latine [Al-Onaizan and Knight, 2002 ; Sherif and Kondrak, 2007 ; Saadane and Semmar, 2012]. La plupart des approches nécessitent une table de correspondance des caractères et des prononciations pour convertir un mot dans une langue source en une séquence de caractères ayant la même prononciation dans la langue cible. Celles-ci différent selon les langues utilisées (langue source/langue cible). La figure 2.1 montre un extrait de la table de translittération des consonnes arabes en caractères latins proposée par ISO 233-2 (1993).

N° d'ordre ISO	Caractère arabe	Code Unicode du caractère arabe	Translittération majuscule	Code Unicode de la translittération majuscule	Translittération minuscule	Code Unicode de la translittération minuscule	Remarques
1	ا	0627					Voir remarques 5.1, 5.2, 5.4 et point 6.4
2	ء	0621	’			02BE	Voir remarque 5.4
3	ب	0628	B	0042	b	0062	
4	ت	062A	T	0054	t	0074	
5	ث	062B	Ṯ	1E6E	ṯ	1E6F	
6	ج	062C	Ǧ	01E6	ǧ	01E7	
7	ح	062D	Ḥ	1E24	ḥ	1E25	
8	خ	062E	Ḫ	0048 puis : 0331	ḫ	1E96	Pour la saisie en majuscule, saisir d'abord « H »
9	د	062F	D	0044	d	0064	
10	ذ	0630	Ḏ	1E0E	ḏ	1E0F	
11	ر	0631	R	0052	r	0072	
12	ز	0632	Z	005A	z	007A	
13	س	0633	S	0053	s	0073	
14	ش	0634	Š	0160	š	0161	
15	ص	0635	Ṣ	1E62	ṣ	1E63	
16	ض	0636	Ḍ	1E0C	ḍ	1E0D	

Fig. 2.1 – Extrait de la table de translittération des consonnes *ISO 233-2*

Des travaux ont proposé des méthodes basées sur la translittération pour reconnaître des entités nommées. Ainsi, Fattah et al. [2006] visent à identifier des noms propres translittérés à partir d'un corpus parallèle anglais-arabe. Dans un premier temps, les noms propres sont extraits dans chaque langue en utilisant le *CLAWS4 POS tagger* pour l'anglais [Leech et al., 1994] et *Buckwalter Arabic Morphological Analyzer Version 1.0* [Buckwalter, 2002] pour l'arabe. Les noms propres arabes sont ensuite romanisés en utilisant la table de romanisation ALA-LC[3]. La figure 2.2 présente un extrait de cette table de romanisation.

ا	ـا	ا	ا	omit (see Note 1)
ب	ـب	بـ	ب	b
ت	ـت	تـ	ت	t
ث	ـث	ثـ	ث	th
ج	ـج	جـ	ج	j
ح	ـح	حـ	ح	ḥ
خ	ـخ	خـ	خ	kh
د	ـد	د	د	d
ذ	ـذ	ذ	ذ	dh
ر	ـر	ر	ر	r
ز	ـز	ز	ز	z
س	ـس	سـ	س	s
ش	ـش	شـ	ش	sh
ص	ـص	صـ	ص	ṣ
ض	ـض	ضـ	ض	ḍ
ط	ـط	طـ	ط	ṭ
ظ	ـظ	ظـ	ظ	ẓ
ع	ـع	عـ	ع	‘ (ayn)
غ	ـغ	غـ	غ	gh
ف	ـف	فـ	ف	f (see Note 2)
ق	ـق	قـ	ق	q (see Note 2)

Fig. 2.2 – Extrait de la table de romanisation ALA-LC

Différentes mesures de similarité ont été utilisées afin de mesurer la similarité entre les noms propres en anglais et ceux translittérés de l'arabe pour ordonner et extraire les bonnes translittérations. Ces mesures sont le coefficient de similarité DICE [Dice, 1945] qui détermine la similarité entre deux échantillons en donnant un indice entre 0 et 1, ainsi que SIM1 et SIM2, deux mesures de similarité proposé par ce travail pour l'arabe. La mesure SIM2 prend en considération l'absence des voyelles courtes dans les textes arabes. L'évaluation atteint une précision et un rappel de 71,4% et 66,5%, respectivement pour, les paires les mieux appariées et une précision et un rappel de 73,8% et 68,2% respectivement pour les trois premières translittérations arabes d'un nom propre anglais. Les résultats montrent également que le système est performant pour les mots avec un faible nombre d'occurrences. De nombreuses translittérations extraites erronées résultent de données insuffisantes.

3. http ://archimedes.fas.harvard.edu/mdh/arabic/arabic-loc.pdf

De même, Semmar and Saadane [2014] proposent un système de translittération des noms propres de l'écriture arabe vers l'écriture latine afin d'améliorer le processus d'alignement des mots simples et composés. La méthode proposée est basée sur des équivalences graphémiques établies à partir d'une étude de corpus de textes parallèles français-arabe. Par exemple :

— La lettre ش est transcrite en S dans DIN-31635, Sh selon UN, EI et ALA-LC, š suivant ISO/R 233 et (ch) dans le corpus d'apprentissage.

— La lettre ظ est transcrite en ẓ dans les différentes normes de translittération et en z, dh et d dans le corpus d'apprentissage.

La première étape consiste à supprimer la dernière voyelle courte ou *tanwin* (marqueur du cas) située à la fin du nom avant de le translittérer. Des règles contextuelles sont alors définies automatiquement à partir de l'outil HTFST, constitué d'une interface basée sur la librairie open-source OpenFst [Riley et al., 2009], pour rendre compte le plus précisément possible des formes observées. Ces règles sont appliquées selon le nombre de consonnes du nom considéré dans un ordre de priorité déterminé. La liste des noms en écriture latine est ensuite normalisée et une pondération est définie comme le nombre d'occurrences retourné par le moteur de recherche Google. Ce système produit en sortie une liste triée de noms arabes translittérés en caractères latins. L'évaluation est effectuée sur 283 phrases du corpus du Monde Diplomatique français-arabe. Les taux de précision, rappel et F-mesure du processus d'alignement augmentent respectivement de 85% , 80% et 82% à 88%, 85% et 86% grâce aux informations produites par le système de translittération.

Selon le type de textes, les tables de translittération peuvent s'avérer peu utile. C'est par exemple le cas lorsqu'il s'agit de *tweet*. Ainsi, Mubarak and Abdelali [2016] présentent une nouvelle approche pour la construction d'un système de translittération. Leur corpus est composé de 881 310 paires de noms de personnes de l'arabe vers l'anglais collectées à partir de Twitter. Les données sont nettoyées par association du nom complet écrit en arabe avec le nom d'utilisateur écrit en caractères latins (sa translittération en anglaise) ainsi que le pays de l'utilisateur. Les noms écrits en arabe et en latin et la localisation des utilisateurs sont normalisés selon la méthode décrite dans Darwish et al. [2012]. Le tableau 2.1 présente des exemples de ces règles de normalisation.

Les lettres arabes qui n'ont pas de correspondance phonétique exacte dans les langues latines sont remplacées par des chiffres remplaçant les caractères arabes. La translittération des noms propres selon les variations dialectales régionales est prise en compte grâce à l'établissement de tables de correspondance prenant en compte les différentes possibilités de translittération des

Caractère initial	Translittération
ي	ى
ة	ه
أ	ا
إ	ا
آ, إ, أ	ا
(>	(A
<	A
|	A

Tab. 2.1 – Extrait de la table de normalisation décrite par Darwish et al. [2012]

caractères arabes en leurs équivalents latins. Pour mesurer et quantifier la similitude entre les noms en arabe et ceux en latin, le score de similarité est calculé en utilisant la distance d'édition de Levenshtein. Les scores obtenus sur 1000 paires de noms sélectionnées aléatoirement sont importants : 0,96 pour la précision, 0,97 pour le rappel et 0,965 pour la F1-mesure.

Les travaux présentés ci-dessus s'intéressent à la translittération des noms propres arabe en anglais, la langue arabe étant considérée comme la langue source. Si notre travail sur la translittération comporte des similarités avec ces travaux, il diffère sur deux points : d'abord, notre objectif est d'extraire la liste des termes arabes à partir d'un corpus parallèle. De plus, nous nous intéressons ici à la translittération inverse : termes anglais translittérés en caractères arabes.

2.4 Conclusion

Dans ce chapitre, nous avons présenté les différents travaux d'acquisition terminologique s'appuyant sur des corpus monolingues et bilingues. Dans nos travaux de recherche, nous nous intéressons à l'acquisition des termes arabes simples et complexes dans un domaine de spécialité. Cependant, les systèmes issus des travaux présentés ci-dessus sur le MSA ne sont pas librement accessibles.

Afin de palier cet inconvénient, nous souhaitons proposer une première approche pour l'extraction de termes en adaptant un système existant. Pour cela, nous avons choisi de nous appuyer sur l'extracteur de termes $\text{Y}_{\text{A}}\text{T}_{\text{E}}\text{A}$ [Aubin and Hamon, 2006] en définissant les règles d'extraction spécifiques au MSA.

Notre deuxième axe de travail consiste à extraire une terminologie arabe à partir d'un corpus bilingue parallèle. Ainsi, nous nous appuyons sur une méthode de transfert translingue tout en tenant compte des spécificités inhérentes à la langue arabe. Nous proposons aussi une méthode

d'extraction des termes arabes translittérés de l'anglais.

Avant de passer à la présentation des différentes stratégies et méthodes que nous proposons pour l'extraction terminologique pour l'arabe standard moderne dans un domaine de spécialité, nous décrivons notre corpus de textes médicaux monolingue arabe et bilingue parallèle dans le chapitre 3.

Chapitre 3

Préparation du corpus

Sommaire

3.1 Introduction

Dans ce chapitre, nous présentons les étapes de collecte, de pré-traitement et d'alignement réalisées pour obtenir notre corpus de travail. Ce corpus doit nous permettre d'atteindre un double objectif : (1) l'adaptation d'un extracteur terminologique existant pour l'arabe standard moderne ; (2) la mise au point d'une méthode d'extraction de termes s'appuyant sur un transfert translingue.

Afin d'effectuer une adaptation d'un extracteur terminologique pour l'arabe standard moderne, nous avons besoin de disposer d'un corpus arabe de spécialité. Cependant, la langue arabe souffre d'une carence de ressources textuelles disponibles. C'est particulièrement le cas en langue de spécialité. À notre connaissance, il n'existe pas encore de corpus composé de textes de spécialité dans la langue arabe disponible librement à des fins d'acquisition terminologique ou de traitement automatique des langues. En effet, même si Al-Sulaiti and Atwell [2006] ont constitué deux corpus de langue générale qui comprennent des sous-corpus de spécialité, les textes utilisés sont issus de deux journaux (Al-Nahar et Al-Hayat) qui sont uniquement accessibles par abonnement. De plus, ces corpus sont composés de textes journalistiques et non des textes produits par des experts. Il ne s'agit donc pas de textes de spécialité.

Par ailleurs, contrairement à la plupart des travaux d'extraction de termes en arabe [Abed et al., 2013], nous ne souhaitons pas travailler sur des données textuelles issues de sites Web ou de forums, car leur qualité terminologique est difficilement vérifiable.

De plus, la mise en œuvre d'une méthode d'extraction de termes basée sur une approche par transfert translingue nécessite de disposer d'un corpus de spécialité parallèle. Il est encore plus difficile de disposer d'un corpus parallèle de textes de spécialité arabe-anglais ou arabe-français et aucun corpus en libre accès ne correspondait à nos besoins. Par conséquent, nous avons été amenés à construire notre propre corpus, un corpus composé d'un ensemble de textes de spécialité parallèles anglais-arabe.

Dans ce contexte, nous avons choisi de constituer un corpus à partir de textes médicaux. En effet, le domaine médical permet d'accéder à de nombreux textes dans différentes langues, en particulier en anglais, en français et en arabe standard moderne. De plus, le domaine médical et de la santé de manière générale sont un domaine dans lequel il existe des terminologies qui pourraient servir de base de départ ou de référence.

Dans la suite de ce chapitre, nous décrivons d'abord les différents traitements mis en œuvre dans la constitution de notre corpus parallèle aligné anglais-arabe et français-arabe dans la section 3.2 : la collecte du corpus, la conversion des documents PDF au format texte, le nettoyage

et la normalisation des textes ainsi que l'étiquetage morpho-syntaxique. Par la suite, dans la section 3.3, nous présentons dans un premier temps l'aligneur GIZA++ et nous détaillons ensuite l'étape d'alignement effectuée en utilisant cet outil.

3.2 Construction du corpus parallèle

Plusieurs travaux s'appuient sur des corpus parallèles [Kontonatsios et al., 2014 ; Hamon and Grabar, 2016 ; Fan et al., 2009], qui sont une ressource importante pour différentes applications comme la traduction automatique, la recherche d'information, etc. De nombreux efforts humains et du temps sont investis dans la collecte de corpus parallèles de textes traduits [Fung and Yee, 1998], et ce type de corpus reste difficile à collecter.

Dans cette section, nous présentons les différentes étapes effectuées afin d'obtenir un corpus parallèle anglais-arabe dans le domaine médical. Nous évoquons les différents problèmes rencontrés lors de l'étape de pré-traitement du corpus. Nous avons jugé qu'il était nécessaire de mettre l'accent sur le processus de nettoyage puisqu'il est important que le corpus soit propre pour les traitements à venir.

Avant tout, nous considérons qu'il est important de présenter la notion de corpus parallèle ainsi que le besoin qui nous a poussé à construire notre propre corpus. Depuis plusieurs décennies, l'extraction de lexiques bilingues d'une manière générale et plus précisément l'acquisition d'une terminologie bilingue, constituent un domaine de recherche important. Les travaux dans ce domaine reposent principalement sur la disponibilité de corpus multilingues composés d'un ensemble de textes dans plusieurs langues. On distingue classiquement deux types de corpus multilingues : les corpus parallèles et les corpus comparables.

Un corpus parallèle est composé d'un ensemble de textes qui présentent des traductions mutuelles. Dans ce contexte, Somers [2001] définit un corpus parallèle comme un ensemble de textes disponibles en deux ou plusieurs langues présentant un texte original et sa traduction, ou d'un texte rédigé par des humains dans plusieurs langues. Selon Fung [1998], un corpus parallèle doit regrouper différents aspects :

1. Les mots ont un seul sens par corpus
2. Les mots ont une traduction unique par corpus
3. Aucune traduction ne manque dans le document cible
4. Les fréquences des occurrences de mots en relation de traduction sont comparables
5. Les positions des mots en relation de traduction sont comparables

Cependant, nous, ainsi que Bouamor [2014], nous n'approuvons pas les deux premiers points imposés par Fung [1998] étant donné qu'il existe des corpus parallèles dont un mot d'une langue source peut correspondre à plusieurs mots en langue cible et/ou avoir différents sens possible comme le corpus des actes du parlement européen *Europarl*.

Un corpus comparable correspond à un ensemble de textes dans différentes langues. Les textes sont issus du même domaine et portent sur le même sujet mais ils ne représentent pas des traductions les unes des autres. Déjean and Éric Gaussier [2002] proposent un critère minimal pour définir les corpus comparables :

> *'Deux corpus de deux langues l1 et l2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue l1, respectivement l2, dont la traduction se trouve dans le corpus de langue l2, respectivement l1.'*

3.2.1 Collecte du corpus

Comme nous l'avons déjà indiqué dans l'introduction, très peu de corpus de spécialité sont disponibles pour la langue arabe [Al-Sulaiti and Atwell, 2006]. C'est particulièrement le cas lorsqu'il s'agit de corpus parallèles. Pour cela, nous avons constitué notre corpus à partir de textes produits par la *United States National Library of Medicine* (NLM), la bibliothèque médicale des Etats-Unis. Ces textes ont été rédigés en langue anglaise et traduits dans de nombreuses autres langues (français, arabe standard moderne, etc.). Les textes collectés sont disponibles en ligne au format PDF, sur le site MedlinePlus[1]. Ces document sont constitués des brochures de quelques pages à destination des patients. Ils fournissent des informations sur des problèmes médicaux, décrivent les conditions de réalisation d'examen, donnent des conseils de comportement face à une maladie, ou pour l'amélioration du bien-être. Ils alternent une page rédigée en anglais et une page rédigée dans une autre langue. Dans la figure 3.1, nous présentons un extrait des documents parallèles anglais-arabe collectés.

3.2.2 Conversion des documents PDF au format texte

Contrairement aux documents en anglais et en français, la conversion au format texte des documents en arabe, nécessaire pour réaliser des traitements automatiques, pose de nombreux problèmes : erreur de forme des caractères, utilisation d'un caractère persan ressemblant graphiquement à un caractère arabe, etc. Ceux-ci rendent les tâches de préparation et de nettoyage

1. `http://www.nlm.nih.gov/medlineplus/languages/all_healthtopics.html`

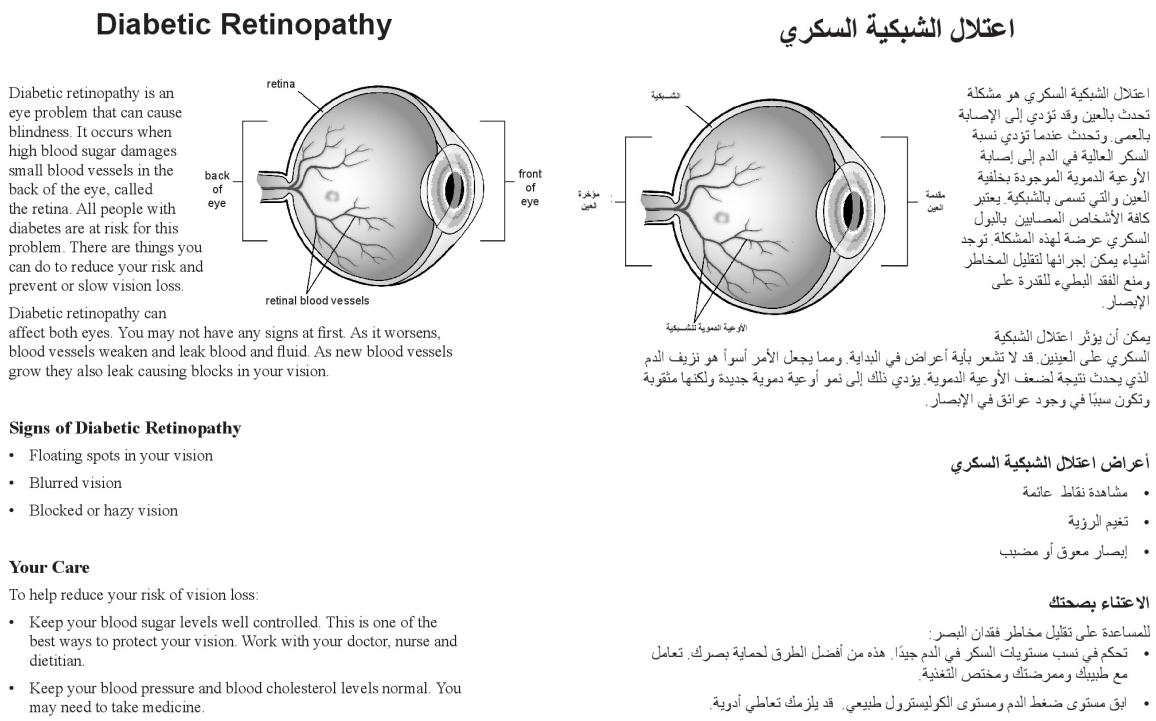


Fig. 3.1 – Extrait des documents parallèles anglais-arabe collectés

du corpus très coûteuse en temps. Nous avons tenté à résoudre ces problèmes en apportant une solution à chacun des problèmes rencontrés (cf. 3.2.3).

Comme le montre la figure 3.1, les documents PDF collectés contiennent des figures porteuses d'informations, notamment sous forme de légendes. Lors du processus de conversion au format texte, nous avons choisi de garder ces informations dont chacune occupe une ligne. La figure 3.2 montre la présence de la documentation dans les fichiers textes résultats.

Actuellement, notre corpus pré-traité est composé de 102 textes parallèles anglais-arabe. Dans le tableau 3.1, nous présentons les caractéristiques de notre corpus

	Nombre de mots
Corpus anglais	35521
Corpus arabe	33402

Tab. 3.1 – Caractéristiques du corpus de travail nettoyé et pré-traité

Tout d'abord, l'étape de conversion a été effectuée grâce à l'utilisation de l'outil `pdftotext` qui est capable de passer, dans un premier temps, du format PDF au format texte. Par la suite, étant donné l'alternance de pages en anglais et en arabe dans le PDF, le fichier texte résultant de la conversion suit également cette structure. Afin d'effectuer des traitements spécifiques à chaque langue, nous avons dû séparer les parties en anglais des parties en arabe à l'aide d'un

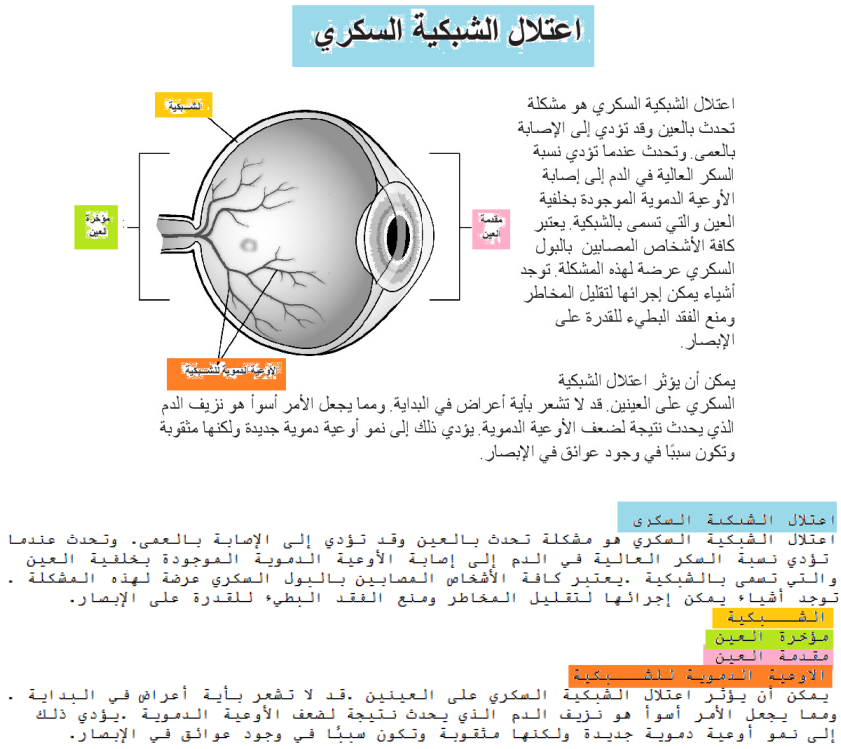


Fig. 3.2 – Prise en compte des informations présentes dans les figures

programme que nous avons écrit spécifiquement pour ces fichiers. A la fin de cette étape, les deux parties sont stockées dans des fichiers séparés.

3.2.3 Nettoyage et pré-traitement

Lors de la conversion au format texte et du nettoyage des documents MedlinePlus en langue arabe, nous avons constaté plusieurs problèmes de codage de caractères dont certains sont déjà mentionnés dans [Habash, 2010]. Ces problèmes entraînent des difficultés de mise en œuvre des approches ou des outils du TAL, comme par exemple, lors de la projection de dictionnaires. Nous avons donc réalisé des traitements spécifiques consistant à nettoyer et à normaliser les textes du corpus. La figure 3.3 présente un extrait non-nettoyé des documents textes produits par la conversion. La figure 3.4 montre le même extrait de notre corpus présenté par celle 3.3 nettoyé parallèle à sa partie anglaise.

Un premier problème concerne les caractères spéciaux UTF-8 utilisés pour le contrôle de texte bidirectionnel (U+202A, U+202B, ...). Ceux-ci sont invisibles dans la majorité des éditeurs de texte et peuvent être introduits de manière irrégulière dans les fichiers PDF ou les fichiers textes obtenus à la suite de l'étape de conversion. Ces caractères provoquent la division des phrases en fragments, chacun étant situé séparément de son contexte (voir la figure 3.3). La

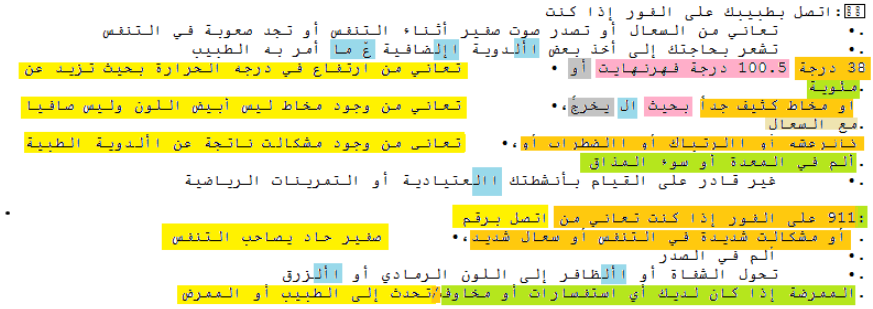


Fig. 3.3 – Extraits non nettoyé d'un document au format texte

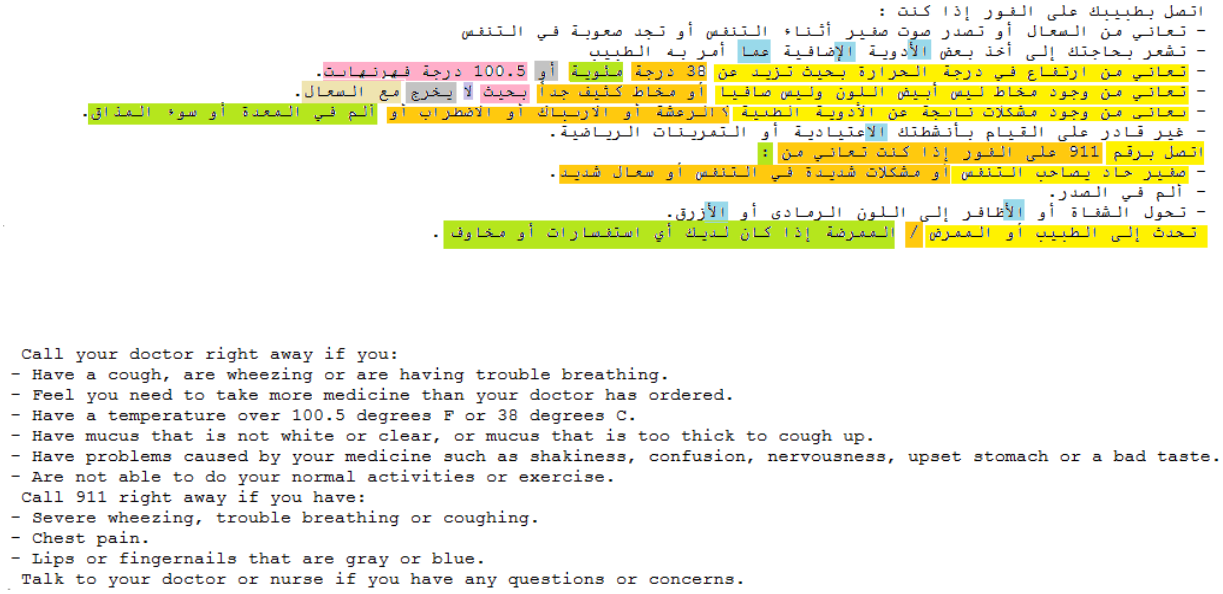


Fig. 3.4 – Texte parallèle arabe-anglais nettoyé

présence des caractères numériques, des symboles de ponctuation ou des caractères latins augmente le risque de présence des caractères de contrôle du texte bidirectionnel. Ces caractères invisibles sont également responsables du désordre produit au niveau de la mise en page lors de la présence des listes à puces. Les traitements basés sur des expressions régulières sont alors perturbés. La suppression de ces caractères de contrôle est donc indispensable. Pour cela, nous avons donc mis au point une étape du nettoyage manuel afin de corriger cette désorganisation et de reconstruire des paragraphes cohérents conformes à la mise en page des documents originaux.

Si ce premier problème est assez simple à corriger, nous avons pu constater d'autres problèmes pouvant empêcher la projection correcte de dictionnaire. Il s'agit d'erreurs liées à la

graphie ou des fautes d'orthographe :

— *Inexactitude de la forme des caractères :*

Dans l'alphabet arabe, quasiment toutes les lettres possèdent plusieurs allographes selon qu'elles se trouvent au début, au milieu ou à la fin d'un mot, ou de manière isolée. Le tableau 3.2 montre les différents allographes de deux lettres arabes[2].

Position	Finale	Médiane	Initiale	Isolée
Allographe de la lettre *Seen*	ـس	ـسـ	سـ	س
Allographe de la lettre *Heh*	ـه	ـهـ	هـ	ه
Allographe de la lettre *Reh*	ـر		ر	

Tab. 3.2 – Variation des allographes des caractères arabes

Dans la plupart de nos fichiers textes, nous avons constaté la substitution des formes correctes de certains caractères par d'autres formes inexactes. Ainsi, ce phénomène apparaît dans le mot **اس‌تنشاقها** pour lequel la forme exacte est **استنشاقها** (*l'inhaler*) ou **امتصاص‌ها** au lieu de **امتصاصها** (*son absorption*). Dans le premier exemple, nous remarquons que le deuxième caractère, la lettre *Seen*--**س**, est présente sous sa forme isolée au lieu d'être sous la forme initiale. Même si elle figure en deuxième position, elle ne peut pas avoir sa forme médiane car elle suit une *Alef*[3] **ا**.

Dans d'autres textes, nous avons observé que la quasi-totalité des caractères sont détachés de ceux qui les précèdent ou les suivent bien qu'ils soient dans leurs formes correctes comme par exemple, dans la phrase suivante :

احرصي على تحديد موعد بأسرع ما يمكن عندما تظنين أنك حاملاً لبدء العناية طوال فترة الحمل قبل الولادة.

Prenez rendez-vous dès que vous pensez être enceinte afin de commencer les soins prénataux.

Ces problèmes cités ci-dessus sont corrigés en passant par une étape de normalisation automatique sur nos textes arabes.

— *Remplacement d'un caractère arabe par un caractère persan qui lui ressemble graphiquement :*

2. Les noms des lettres arabes utilisés sont ceux proposés par Unicode.
3. La lettre *Alef* n'apparaît jamais attachée à ce qui la suit.

Il est, par exemple, possible de rencontrer le caractère farsi ی (U+06CC, *Arabic Letter Yeh Farsi*) à la place du ى (U+0649, *Alef Maksura*)[4]. La forme utilisant le caractère initial farsi **یقوم** et la forme utilisant caractère initial arabe **يقوم** (*se base*) sont similaires graphiquement lorsque les caractères sont détachés (position isolée), mais lorsqu'il est dans sa forme attachée, le caractère arabe apparaît différemment. Il est possible alors, de retrouver la forme farsi, erronée, dans des textes arabes. Nous tenons à mentionner que ce problème se rencontre aussi dans d'autres langues. Par exemple, le caractère « p » peut correspondre au « p » latin ou le « r » cyrillique.

Pour ce problème de remplacement d'un caractère arabe par un caractère persan, nous avons procédé à une étape de normalisation automatique des textes arabes pour les corriger.

— *Remplacement d'un caractère arabe par un autre caractère arabe :*

Ainsi, nous remarquons que, dans certains textes, le caractère **ك** est remplacé par آ dans le mot **مشاآل**, alors que la forme correcte du mot est **مشاكل** (*problèmes*). L'exemple (3) présente une comparaison entre l'ensemble des caractères de ces deux mots. Il en est de même pour le mot **كيس** (*sac/poche*) alors que la forme correcte du mot est **آيس**.

(3) mot erroné : م ش ا آ ل
Mot correct : م ش ا ك ل

Pour corriger ce problème, nous avons vérifié manuellement si l'occurrence du caractère آ est le résultat d'une mauvaise formation des mots lors de la conversion au format texte, ou si elle représente le caractère initial du mot en question comme par exemple dans **الآثار** (*les effets*) ou dans **آمنة** (*en sécurité*). La correction des caractères erronés ne peut être effectuée que manuellement.

— *Inversion de la position de caractères :*

Lors de la phase de nettoyage, nous avons également rencontré des inversions de caractères à l'intérieur d'un mot. Généralement, cette inversion se limite à la permutation des positions des caractères ا -- ***A*** (la lettre *alif*) et ل -- ***L*** (la lettre *lem*). Ce problème se manifeste dans la plupart des textes arabes de notre corpus lorsque le caractère *alif* précède le caractère *lem*. Par exemple, il est possible de rencontrer la forme (**ألوعية** au lieu de **الأوعية** (les vaisseaux) (voir 4) ou **الخاليا** au lieu de **الخلايا** (les cellules).

4. La similitude et la différence graphique sont liées à la police de caractères utilisée.

(4) الأوعية

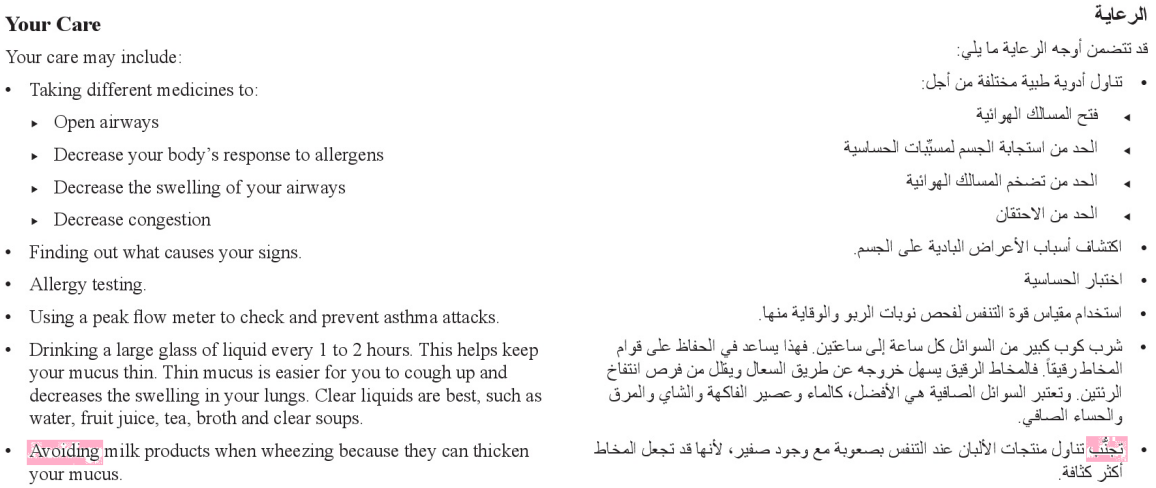
Ordre initial des caractères : ا ل أ و ع ي ة
Ordre inversé des caractères : ا أ ل و ع ي ة

Afin de corriger ces problèmes, nous avons effectué une vérification manuelle car la détection de ces inversions exige des connaissances en littérature arabe pour distinguer les couples de caractères en bonne position de celles erronées qui seront par la suite corrigées manuellement.

Bien que notre corpus soit composé d'un ensemble de textes non voyellés, quelques voyelles courtes peuvent être disposées sur certains caractères comme dans **تُنتِج** (*elle produit*). Dans la plupart des cas, nous faisons l'hypothèse que les rédacteurs ont placé des voyelles pour lever une ambiguïté et que les lecteurs interprètent correctement le sens qu'ils souhaitent présenter. Par exemple, dans un même texte, nous avons trouvé le mot **تجنب** sans voyelle et avec deux voyellations différentes : **تجنُّب** (*l'évitement*) et **تجنَّب** (*il a évité/évite*). Ici, les deux voyellations sont correctes (la phrase est lisible et correcte avec chacune des deux voyellations) mais il est probable que le rédacteur ait voulu préciser qu'il s'agit d'un verbe ou d'une forme déverbale. Même si dans le deuxième mot il existe encore une ambiguïté, celle-ci peut être résolue par le contexte car il s'agit d'une liste de traitements conseillés pour les patients (voir figure 3.5). Dans ce cas, il suffit de préciser qu'il s'agit d'un verbe pour désambiguïser le mot. Dans le même texte, le mot **تجنب** est présent dans sa forme non voyellée car il figure dans la même liste des traitements. Le lecteur peut directement voir qu'il s'agit d'un verbe à l'impératif (voir figure 3.6).

Si la présence de voyelles aide le lecteur, elle peut être la cause de problème lors de la conversion en format texte. Parfois, elles figurent isolées avant ou après au caractère sur lequel elles interviennent comme dans **حَّكة**. Dans cet exemple, la première voyelle *Fatha* /a/ ◌َ intervient sur le premier caractère **ح** alors que la deuxième *Shadda* ◌ّ est associée au deuxième caractère **ك**. La forme correcte du mot est **حكّة** (*démangeaison*).

Une partie de ces erreurs sont corrigées automatiquement lors de l'étape de nettoyage et normalisation des textes. Cependant, la vérification manuelle est nécessaire étant donné l'incohérence des problèmes rencontrés.



Your Care

Your care may include:

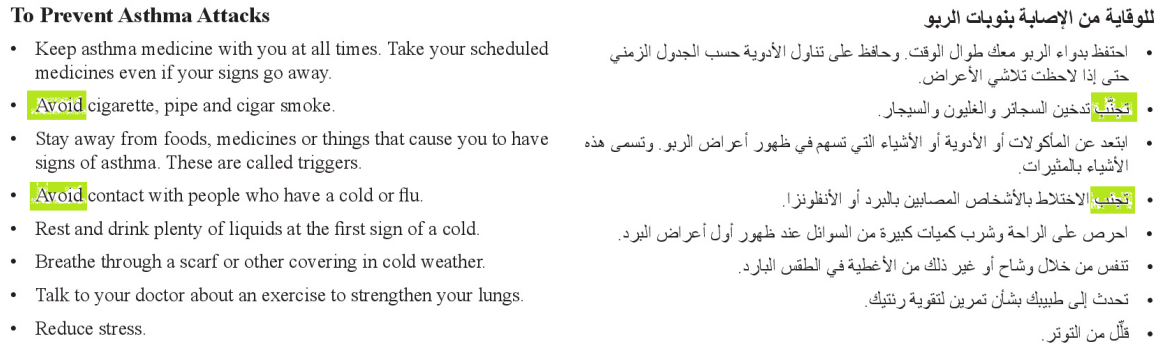
- Taking different medicines to:
  - Open airways
  - Decrease your body's response to allergens
  - Decrease the swelling of your airways
  - Decrease congestion
- Finding out what causes your signs.
- Allergy testing.
- Using a peak flow meter to check and prevent asthma attacks.
- Drinking a large glass of liquid every 1 to 2 hours. This helps keep your mucus thin. Thin mucus is easier for you to cough up and decreases the swelling in your lungs. Clear liquids are best, such as water, fruit juice, tea, broth and clear soups.
- Avoiding milk products when wheezing because they can thicken your mucus.

الرعاية

قد تتضمن أوجه الرعاية ما يلي:

- تناول أدوية طبية مختلفة من أجل:
  - فتح المسالك الهوائية
  - الحد من استجابة الجسم لمسبّبات الحساسية
  - الحد من تضخم المسالك الهوائية
  - الحد من الاحتقان
- اكتشاف أسباب الأعراض البادية على الجسم.
- اختبار الحساسية
- استخدام مقياس قوة التنفس لفحص نوبات الربو والوقاية منها.
- شرب كوب كبير من السوائل كل ساعة إلى ساعتين. فهذا يساعد في الحفاظ على قوام المخاط رقيقاً. فالمخاط الرقيق يسهل خروجه عن طريق السعال ويقلل من فرص انتفاخ الرئتين. وتعتبر السوائل الصافية هي الأفضل، كالماء وعصير الفاكهة والشاي والمرق والحساء الصافي.
- تجنّب تناول منتجات الألبان عند التنفس بصعوبة مع وجود صفير، لأنها قد تجعل المخاط أكثر كثافة.

Fig. 3.5 – Utilisation de la forme déverbale du mot arabe non-voyellé **تجنب**



To Prevent Asthma Attacks

- Keep asthma medicine with you at all times. Take your scheduled medicines even if your signs go away.
- Avoid cigarette, pipe and cigar smoke.
- Stay away from foods, medicines or things that cause you to have signs of asthma. These are called triggers.
- Avoid contact with people who have a cold or flu.
- Rest and drink plenty of liquids at the first sign of a cold.
- Breathe through a scarf or other covering in cold weather.
- Talk to your doctor about an exercise to strengthen your lungs.
- Reduce stress.

للوقاية من الإصابة بنوبات الربو

- احتفظ بدواء الربو معك طوال الوقت. وحافظ على تناول الأدوية حسب الجدول الزمني حتى إذا لاحظت تلاشي الأعراض.
- تجنّب تدخين السجائر والغليون والسيجار.
- ابتعد عن المأكولات أو الأدوية أو الأشياء التي تسهم في ظهور أعراض الربو. وتسمى هذه الأشياء بالمثيرات.
- تجنب الاختلاط بالأشخاص المصابين بالبرد أو الأنفلونزا.
- احرص على الراحة وشرب كميات كبيرة من السوائل عند ظهور أول أعراض البرد.
- تنفس من خلال وشاح أو غير ذلك من الأغطية في الطقس البارد.
- تحدث إلى طبيبك بشأن تمرين لتقوية رئتيك.
- قلّل من التوتر.

Fig. 3.6 – Utilisation de la forme verbale du mot arabe non-voyellé **تجنب**

3.2.4 Analyse morphologique et étiquetage morpho-syntaxique

Comme nous l'avons déjà indiqué dans l'introduction, nous proposons, dans un premier temps, une méthode d'extraction des termes consistant en l'adaptation à l'arabe standard moderne de l'extracteur terminologique YaTeA [Aubin and Hamon, 2006]. Puisque YaTeA utilise des textes étiquetés morpho-syntaxiquement et lemmatisés pour extraire des termes candidats, nous avons effectué une analyse morphologique de nos textes médicaux en arabe. L'étiquetage morpho-syntaxique est une étape du TAL qui permet d'associer à chaque mot du texte sa catégorie grammaticale correspondante.

Pour ce faire, nous avons envisagé l'utilisation des deux outils : Stanford Tagger [Toutanova et al., 2003] et MADA+TOKAN [Habash et al., 2010]. Les précédentes comparaisons de ces outils [Green and Manning, 2010 ; Albogamy and Ramsay, 2015] montrent que les résultats obtenus avec MADA+TOKAN sont de meilleure qualité. Ceci est confirmé par les tests faits sur un échantillon. Nous avons donc opté pour l'outil MADA+TOKAN qui, outre la qualité de

son analyse morphologique, offre une lemmatisation des mots du corpus.

Dans un deuxième temps, nous avons aussi utilisé l'outil MADAMIRA[5] [Pasha et al., 2014]. pour se servir des différents fichiers résultat obtenus avec des textes non agglutinés. Ceux-ci seront utilisés pour améliorer l'alignement des corpus anglais-arabe. MADAMIRA est un système d'analyse morphologique et de désambiguïsation de l'arabe qui combine les aspects de deux systèmes de traitement de l'arabe qui sont MADA [Habash et al., 2009] et AMIRA [Diab et al., 2007]. Pour chaque mot arabe, MADA produit une liste d'analyses décrivant toutes les interprétations morphologiques possibles de ce mot comme la diacritisation, l'étiquette grammaticale, le lemme et les caractéristiques flexionnelles et clitiques. Le système applique ensuite un ensemble de modèles SVM *(Support Vector Machines)* et des modèles de langue *n-grammes* afin de produire les prédictions par mot. Quant à AMIRA, l'outil intègre un segmenteur, un étiqueteur morphologique et un analyseur syntaxique de surface. Il repose sur un apprentissage supervisé sans dépendance explicite de la connaissance de la morphologie profonde [Pasha et al., 2014].

MADA+TOKAN est un outil polyvalent, personnalisable, dont les informations en sortie peuvent être choisies. Il est aussi disponible en ligne pour être testé sur une ou deux phrases. Cet outil est constitué de deux composants : MADA et TOKEN. MADA associe à chaque mot d'une phrase en arabe une description lexicale et morphologique : sa catégorie grammaticale, ses traits morpho-syntaxiques, son lemme, son lexème, sa diacritisation et son analyse morphologique. Par la suite, tout en tenant compte des informations produites par MADA, TOKAN génère une tokenisation (une segmentation) formatée selon les spécifications de l'utilisateur. Cette tokenisation identifie également la racine du mot.

La version que nous utilisons, MADA 3.0, comprend un pré-traitement qui convertit le texte en encodage Buckwalter pour les traitements internes. Les résultats produits sont aussi présentés en Buckwalter. MADA+TOKAN fournit également, plusieurs traits morphologiques associés aux catégories grammaticales (nom, verbe, adverbe...) des mots. La figure 3.7 présente un extrait de l'analyse morpho-syntaxique du mot الدم -- ***Aldm*** (*le sang*), produite par MADA+TOKAN (la signification des différents caractéristiques associées au mot est donnée ci-dessous).

Parmi toutes ces caractéristiques, nous avons extrait les traits morphologiques que nous jugeons nécessaires pour notre travail. La liste suivante présente ces traits morphologiques ainsi que leurs valeurs associées :

— le genre : masculin, féminin, non applicable, non défini

5. `https://camel.abudhabi.nyu.edu/madamira/?ref=all`

```
;;WORD Aldm
;;SVM_PREDICTIONS : Aldm asp :na cas :g enc0 :0 gen :m mod :na num :s per :na
pos :noun prc0 :Al_det prc1 :0 prc2 :0 prc3 :0 stt :d vox :na
1.019825 diac :Ald ami lex :dam_1 bw :Al/DET+dam/NOUN+i/CASE_DEF_GEN
gloss :blood pos :noun prc3 :0 prc2 :0 prc1 :0 prc0 :Al_det per :na asp :na
vox :na mod :na gen :m num :s stt :d cas :g enc0 :0 rat :y source :lex stem :dam
stemcat :N
```

Fig. 3.7 – Extrait de l'analyse morpho-syntaxique du mot **الدم** – ***Aldm*** (*le sang*), produite par MADA+TOKAN

— le nombre : singulier, dual, pluriel, non applicable, non défini

— le cas : nominatif -- le mot portera à la fin une damma ou double damma, accusatif -- le mot portera à la fin une fatha ou double fatha, génitif --le mot portera à la fin une kasra ou double kasra, non applicable, non défini

— l'état : indéfini, défini, possesseur/*idafa*, non applicable, non défini

La figure 3.3 présente un aperçu des caractéristiques pertinentes extraites du résultat de l'étiquetage morpho-syntaxique de la phrase présentée dans l'exemple (5) contenant le mot **الدم** -- ***Aldm*** (*le sang*).

(5) Arabe : يتم إدخال الدم عبر شريان وريدي إلى الأوعية الدموية
Buckwalter : Aldmwyp AlAwEyp Aly wrydy $ryAn Ebr Aldm AdxAl ytm
Français : sanguins vaisseaux aux veineux artère par le-sang entrer est-fait

Le sang est inséré à travers l'artère veineuse dans les vaisseaux sanguins

ytm	verb	m	s	na	na	tam -i
AdxAl	noun	m	s	n	c	<idoxAl
Aldm	noun	m	s	g	d	dam
Ebr	noun	m	s	a	c	Eabor
$ryAn	noun	m	s	g	c	$iroyAn
wrydy	adj	m	d	g	c	wariyd
Aly	prep	na	na	na	na	<ilaY
AlAwEyp	noun	f	s	g	d	wiEA’
Aldmwyp	adj	f	s	g	d	damawiy

Tab. 3.3 – Caractéristiques pertinentes de l'étiquetage morpho-syntaxique d'une phrase arabe en Buckwalter

A partir du résultat de MADA+TOKAN, nous avons créé notre propre fichier de sortie afin

de permettre une utilisation aisée par YATEA. Le format de ce fichier est identique au fichier résultat du TreeTagger :

Forme-Fléchie POS Forme-Lemmatisée

Pour chaque mot, nous avons créé une étiquette construite de la manière suivante : catégorie-genre-nombre-cas-état. Pour assurer le bon fonctionnement de YATEA adapté au MSA, nous avons converti les formes fléchies et lemmatisées des mots arabes au format Buckwalter en caractères arabes à l'aide du module Perl `Unicode ::Normalize`. La figure 3.4 présente un exemple de fichier de sortie du post-traitement après l'étape d'étiquetage morpho-syntaxique.

Forme-Fléchie	POS	Forme-Lemmatisée	Glose
يتم	verb-m-s-na-na	**تَمّ**	***se fait***
ادخال	noun-m-s-n-c	**إِدْخال**	***entrer***
الدم	noun-m-s-g-d	**دَم**	***le sang***
عبر	noun-m-s-a-c	**عَبْر**	***par***
شريان	noun-m-s-g-c	**شِرْيان**	***Artère***
وريدي	adj-m-d-g-c	**وَرِيد**	***veineuse***
الي	prep-na-na-na-na	**إِلَى**	***à/au/aux***
الاوعية	noun-f-s-g-d	**وِعاء**	***les vaisseaux***
الدموية	adj-f-s-g-d	**دَمَوِيّ**	***sanguins***

Tab. 3.4 – Extrait du fichier de sortie suite à l'étape d'étiquetage morpho-syntaxique

3.3 Alignement des textes au niveau des mots

Afin de construire une terminologie arabe en nous appuyant sur une méthode de transfert translingue, il est nécessaire de réaliser une étape d'alignement en préliminaire au processus d'extraction des termes arabes. L'alignement consiste à mettre en correspondance les différentes unités linguistiques des textes de deux langues différentes, ayant une relation de traduction. Il existe différents niveaux de granularité pour effectuer l'alignement de deux textes : le niveau des paragraphes, des phrases, des segments ou des mots. Ce niveau sera déterminé en fonction des besoins et de l'application. De plus, les résultats d'un niveau dépendent des niveaux plus élevés. Par exemple, la qualité d'alignement au niveau des mots peut dépendre de la qualité d'alignement au niveau des phrases.

Dans le cadre de l'acquisition terminologique multilingue à partir d'un corpus parallèle, un alignement au niveau des mots peut être réalisé en utilisant l'outil GIZA++ [Och and Ney, 2003]. L'alignement est utilisé pour mettre en correspondance les termes extraits des deux langues [Fan

et al., 2009 ; Ayed et al., 2012] ou d'extraire l'équivalence de traduction des termes dans une langue cible [Kontonatsios et al., 2014 ; Hamon and Grabar, 2016]. D'autres travaux ont effectué un alignement au niveau des séquences en utilisant la boite à outil Moses pour la traduction automatique [Koehn et al., 2007], pour l'extraction automatique des termes bilingues à partir de corpus parallèles [Ideue et al., 2011] ou même au niveau des documents à partir de corpus comparables pour l'extraction de lexique bilingue [Vulic and Moens, 2015].

3.3.1 Processus d'alignement

Pour l'ensemble des documents de notre corpus, les textes ont été segmentés et alignés manuellement au niveau des paragraphes au fur et à mesure de l'étape de nettoyage et pré-traitement du corpus présenté ci-dessus dans la partie 3.2.3. Pour les informations extraites à partir des figures (voir fig. 3.1), nous avons bien placé chaque annotation sur une ligne en gardant le même ordre dans les textes anglais et arabes.

Par la suite, un alignement au niveau des phrases a été effectué automatiquement, dans un premier temps, en nous basant sur les signes de ponctuation. Par la suite, nous avons vérifié la qualité de cet alignement manuellement. Il est important de noter que dans certains cas, une phrase de la langue L_1 est alignée avec deux phrases ou plus de la langue L_2. Autrement dit, une phrase en MSA peut être équivalente à deux phrases en anglais. Ce déséquilibre au niveau des phrases provient de la diversité des structures linguistiques entre la langue arabe et la langue anglaise mais aussi, parfois à des omissions au moment de la traduction. La figure 3.8 présente un extrait d'un paragraphe aligné anglais-arabe dont une phrase longue en arabe peut correspondre à deux phrases en anglais. Chaque couleur correspond une phrase en arabe alignée avec deux phrases en anglais.

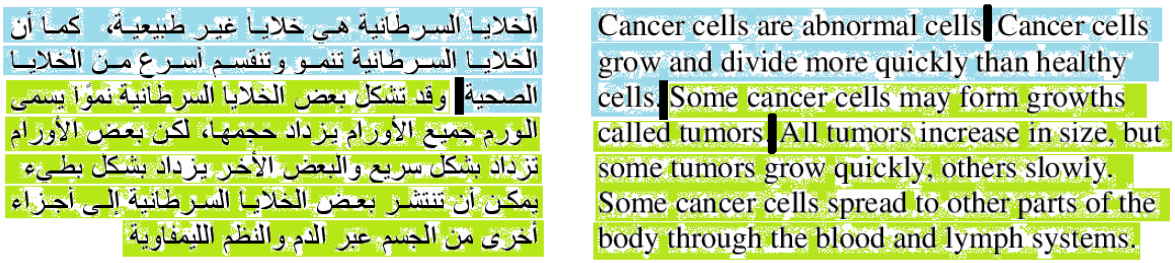


Fig. 3.8 – Extrait d'un paragraphe anglais-arabe présentant les différences de répartition des phrases

Nous avons constaté aussi que l'absence des signes de ponctuation à la fin de certains paragraphes perturbe l'équivalence et l'équilibre entre les différents segments de textes dans les deux

langues (cf. la fin du paragraphe en arabe, dans la figure 3.8). Afin de mener à bien notre alignement au niveau des phrases, nous avons choisi de diviser manuellement ces phrases longues en deux phrases distinctes en ajoutant des points comme signe de fin des phrases tout en respectant la segmentation définie pour la partie en anglais. Parfois, nous procédons aussi à l'élimination de quelques conjonctions de coordination pour assurer et garder la pertinence de la structure des phrases sans toucher à la sémantique du contenu. La figure 3.9 présente la version texte du paragraphe arabe présenté à la figure 3.8.

الخلايا السرطانية هي خلايا غير طبيعية .
كما أن الخلايا السرطانية تنمو وتنقسم أسرع من الخلايا الصحية.
وقد تشكل بعض الخلايا السرطانية نموا يسمى الورم.
جميع الأورام يزداد حجمها ،لكن بعض الأورام تزداد بشكل سريع والبعض الأخر يزداد بشكل بطيء.
يمكن أن تنتشر بعض الخلايا السرطانية إلى أجزاء أخرى من الجسم عبر الدم والنظم الليمفاوية.

```
Cancer cells are abnormal cells.
Cancer cells grow and divide more quickly than healthy cells.
Some cancer cells may form growths called tumors.
All tumors increase in size, but some tumors grow quickly, others slowly.
Some cancer cells spread to other parts of the body through the blood and lymph systems.
```

Fig. 3.9 – Correction de la structure du corpus arabe pour l'alignement

Avant de passer à la dernière étape d'alignement, nous tenons à préciser qu'à ce stade d'avancement, notre corpus est plutôt aligné au niveau des lignes qu'au niveau des phrases. Autrement dit, une ligne ne représente pas forcément une phrase. Elle peut aussi comporter un titre (titre du document, titre de section, titre de sous-section, etc.), la documentation qui présente dans les figures, un élément d'une liste à puce, etc. L'alignement réalisé respecte à la fois la structure du document et la segmentation en phrases. La figure 3.10 montre le résultat final de l'alignement de l'extrait des documents présenté à la figure 3.1.

La dernière étape d'alignement consiste à réaliser un alignement mot à mot à l'intérieur des paires de phrases alignées. Nous nous sommes alors servis de l'outil GIZA++ [Och and Ney, 2003]. Depuis plusieurs années, cet outils est reconnu comme la référence pour effectuer un alignement des mots. Il repose sur les modèles probabilistes Des modèles statistiques d'alignement de mots sont calculés avec une complexité croissante en construisant des liens bidirectionnels entre les mots d'une phrase dans les deux langues (langue source et langue cible), puis les résultats sont symétrisés. À l'exécution de GIZA++, la langue source est d'abord projetée dans la langue cible puis inversement, la langue cible est projetée dans la langue source. La sortie de GIZA++ est post-traitée en utilisant les trois heuristiques de symétrisation décrites dans [Och and Ney, 2003].

اعتلال الشبكية السكري
اعتلال الشبكية السكري هو مشكلة تحدث بالعين وقد تؤدي إلى الإصابة بالعمى .
وتحدث عندما تؤدي نسبة السكر العالية في الدم إلى إصابة الأوعية الدموية الموجودة بخلفية العين والتي تسمى بالشبكية .
يعتبر كافة الأشخاص المصابين بالبول السكري عرضة لهذه المشكلة .
توجد أشياء يمكن إجرائها لتقليل المخاطر ومنع الفقد البطيء للقدرة على الإبصار .
الشـــبكية
مؤخرة العين
مقدمة العين
الأوعية الدموية للشـــبكية
يمكن أن يؤثر اعتلال الشبكية السكري على العينين .
قد لا تشعر بأية أعراض في البداية .
ومما يجعل الأمر أسوأ هو نزيف الدم الذي يحدث نتيجة لضعف الأوعية الدموية .
يؤدي ذلك إلى نمو أوعية دموية جديدة ولكنها مثقوبة وتكون سببًا في وجود عوائق في الإبصار .
أعراض اعتلال الشبكية السكري
- مشاهدة نقاط عائمة
- تغيم الرؤية
- إبصار معوق أو مضبب
الاعتناء بصحتك
للمساعدة على تقليل مخاطر فقدان البصر :
- تحكم في نسب مستويات السكر في الدم جيدًا .
هذه من أفضل الطرق لحماية بصرك .
تعامل مع طبيبك وممرضتك ومختص التغذية .
- ابق مستوى ضغط الدم ومستوى الكوليسترول طبيعي .

diabetic retinopathy
diabetic retinopathy is an eye problem that can cause blindness .
it occurs when high blood sugar damages small blood vessels in the back of back of the eye , called eye the retina .
all people with diabetes are at risk for this problem .
there are things you can do to reduce your risk and prevent or slow vision loss .
 retina
 back of eye
 front of eye
 retinal blood vessels
 diabetic retinopathy can affect both eyes .
you may not have any signs at first .
as it worsens , blood vessels weaken and leak blood and fluid .
as new blood vessels grow they also leak causing blocks in your vision .
 signs of diabetic retinopathy
- floating spots in your vision
- blurred vision
- blocked or hazy vision
 your care
 to help reduce your risk of vision loss :
- keep your blood sugar levels well controlled .
this is one of the best ways to protect your vision .
work with your doctor , nurse and dietitian .
- keep your blood pressure and blood cholesterol levels normal .
you may need to take medicine .

Fig. 3.10 – Résultat final de l’alignement

3.3.2 Amélioration de la qualité d’alignement

Un mot arabe est souvent caractérisé par sa morphologie concaténative (1.3.2.3). Autrement dit, la plupart des mots arabes sont agglutinés. Cette caractéristique influence le déroulement et les performances des outils du traitement automatique du MSA.

Ainsi, la qualité des ressources bilingues extraites à partir de corpus parallèles dépend principalement de la qualité de l’alignement [Chiao, 2004]. Pour cette raison, nous avons testé différents alignements, présentés ci-dessous tout en étudiant l'impact de l'agglutination sur l'alignement des textes arabes au niveau des mots.

Dans ce contexte, nous avons utilisé l'outil MADAMIRA. Outre l'étiquetage morpho-syntaxique, MADAMIRA offre d'autres nouveaux modèles de fichiers. Il produit les textes

arabes de notre corpus sous deux différentes formes non agglutinées dont nous nous servons afin d'améliorer la qualité d'alignement de notre corpus anglais-arabe. Le premier fichier contient le corpus arabe désagglutiné au niveau des clitiques en gardant toujours les articles attachés aux mots qu'ils définissent. Le deuxième fichier contient nos textes arabes fournis en entrée en séparant tous les proclitiques et les enclitiques des mots auxquels ils sont liés, y compris les articles définis.

Afin d'obtenir une meilleure qualité des résultats, nous avons effectué trois différents alignements au niveau des mots entre le corpus anglais ainsi que les différents motifs de texte du corpus arabe désagglutiné.

— Alignement1 : Alignement du corpus anglais et arabe réalisé sans tenir compte de la désagglutination : les proclitiques et les enclitiques sont agglutinés au mot auquel ils se rapportent.

— Alignement2 : Alignement du corpus anglais et arabe en tenant compte de la désagglutination des clitiques sauf des articles : les enclitiques et les proclitiques sont désagglutinés du mot auquel ils se rapportent. La segmentation morphologique est réalisée par MADAMIRA.

— Alignement3 : Alignement du corpus anglais et arabe en tenant compte de la désagglutination des clitiques : les enclitiques, les proclitiques et les articles sont désagglutinés du mot auquel ils se rapportent. La segmentation morphologique est réalisée par MADAMIRA.

Le tableau 3.5 présente les différentes caractéristiques des trois corpus ci-dessus avant et après sélection des alignements ayant les probabilités les plus élevées pour chaque mot anglais. Les résultats obtenus seront présentés par la suite à la section 5.2 du chapitre 5.

	Alignement1		Alignement2		Alignement3	
	Avant sélection	Après sélection	Avant sélection	Après sélection	Avant sélection	Après sélection
Nombre de couple de mots	27800	3917	20235	3917	18190	3917
Nombre de mots anglais	3917	3917	3917	3917	3917	3917
Nombre de mots arabes	9669	3342	6435	3071	5395	2848

Tab. 3.5 – Caractéristiques des corpus d'alignement avant et après sélection

3.4 Conclusion

Dans ce chapitre, nous avons présenté les différentes étapes de construction de notre corpus de spécialité parallèle anglais-arabe. Le corpus est composé d'un ensemble de textes médicaux issus de brochures à destination des patients. Nous avons exposé les différents problèmes liés à la langue arabe que nous avons rencontrés lors de l'étape semi-automatique de nettoyage et de pré-traitement du corpus. Nous avons ensuite utilisé l'outil MADA+TOKAN pour effectuer l'étape d'étiquetage morpho-syntaxique et la lemmatisation du corpus arabe.

Afin de réaliser un alignement au niveau des mots entre l'ensemble de nos textes bilingues parallèles anglais-arabe, nous avons eu recours à l'outil GIZA++. De plus, nous avons utilisé les différents modèles de textes désagglutinés produits par l'étiquetage morpho-syntaxique MADAMIRA. Notre objectif était d'améliorer la qualité d'alignement de notre corpus anglais-arabe en agissant sur l'une des caractéristiques du MSA (voir la section 1.3.2.3 du chapitre 1).

Après avoir construit notre corpus parallèle et aligné anglais-arabe, nous passons maintenant aux différentes méthodes d'acquisition terminologique que nous avons proposées pour le MSA. Nous présentons ainsi au chapitre 4 trois stratégies pour l'extraction de termes médicaux arabes à partir de notre corpus.

Chapitre 4

Extraction terminologique pour l'arabe

Sommaire

4.1 Introduction

Dans ce chapitre, nous proposons plusieurs méthodes pour l'acquisition terminologique en arabe standard moderne (MSA). Nous nous intéressons à l'extraction de termes candidats à partir d'un corpus de spécialité. Ces méthodes visent à lever les verrous que constitue la faible disponibilité des ressources et des outils terminologiques pour la langue arabe. Pour cela, nous nous appuyons sur plusieurs outils linguistiques et terminologiques en arabe, anglais et français.

Nous commençons par présenter une adaptation d'un extracteur terminologique de la langue française et anglaise vers le MSA (section 4.2). Puis, nous proposons une méthode qui permet d'extraire des termes médicaux anglais-arabe produits à partir de la translittération des termes anglais en caractères arabes (section 4.3). Finalement, à la section 4.4, nous définissons une méthode d'acquisition terminologique pour le MSA basée sur une méthode de transfert translingue en s'appuyant sur un corpus parallèle anglais-arabe dans le domaine médical.

4.2 Adaptation de YaTeA pour l'arabe

Afin de construire des ressources terminologiques en MSA, nous proposons, dans un premier temps, une approche permettant d'identifier des termes simples et complexes dans des textes de spécialité arabes en utilisant les outils existants pour d'autres langues. Nous proposons d'adapter un extracteur terminologique offrant la possibilité d'intégrer des spécificités de la langue arabe dans le processus d'extraction de termes. Pour ce faire, nous nous servons de l'outil librement disponible YaTeA[1] [Aubin and Hamon, 2006].

L'adaptation a d'abord consisté à décrire le processus d'extraction des termes arabes de manière similaire au processus déjà défini pour l'anglais et le français tout en prenant en compte quelques particularités morpho-syntaxiques de la langue arabe comme l'agglutination et la non voyellation.

4.2.1 L'extracteur terminologique YaTeA

L'extracteur terminologique YaTeA offre la possibilité de prendre en compte les particularités d'une langue, ici le MSA, grâce à la définition de règles d'analyse superficielle et des ressources linguistiques. YaTeA ayant été d'abord développé pour analyser des textes en français et en anglais, l'adaptation de l'approche à une langue sémitique est un défi. Cela nous permettra aussi

1. http://search.cpan.org/~thhamon/Lingua-YaTeA/

de connaître les limites de la méthode implémentée dans YATEA.

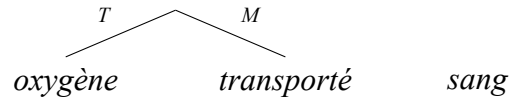
Cet extracteur terminologique identifie les expressions nominales pouvant correspondent aux termes d'un domaine. Le processus d'extraction des termes réalise une analyse syntaxique superficielle de textes étiquetés morpho-syntaxiquement et lemmatisés en plusieurs étapes [Hamon et al., 2014] :

Il y a moins d'oxygène transporté dans le sang.

(a) Texte initial

oxygène transporté sang

(b) Étape 1 : Découpage à l'aide de frontières syntaxiques



(c) Étape 2 : Analyse syntaxique des termes candidats

Fig. 4.1 – Etapes d'extraction des termes candidats sur une phrase en français

— **Etape 1** : Le texte est découpé à l'aide de frontières syntaxiques positives et négatives. Ces frontières peuvent être des étiquettes morpho-syntaxiques (pronoms, verbes conjugués, marques typographiques, etc.) mais aussi des mots ou des groupes de mots. Cette première étape permet d'obtenir des syntagmes nominaux maximaux pouvant constituer ou contenir des termes candidats.

  Ainsi, l'application des frontières syntaxiques sur l'exemple 4.1a de la figure 4.1 permet d'identifier les syntagmes maximaux *oxygène transporté* et *sang* (figure 4.1b). Une série de post-traitements permet d'améliorer la qualité des syntagmes nominaux maximaux à analyser syntaxiquement (catégories morpho-syntaxiques ne pouvant être présentes au début ou à la fin des syntagmes, séquences de mots comme les locutions prépositionnelles, ou d'étiquettes morpho-syntaxiques conduisant inévitablement à une analyse syntaxique erronée).

— **Etape 2** : Des patrons d'analyse syntaxique prenant en compte des informations morpho-syntaxiques des termes sont appliqués récursivement. Cela conduit à identifier les syntagmes correspondant aux termes candidats. Chaque terme candidat est représenté sous la forme d'un arbre binaire de constituants ayant soit le rôle de tête soit le rôle de modifieur. Les constituants sont également considérés comme des termes candidats. Les syntagmes

pour lesquels il n'est pas possible de produire une analyse syntaxique sont considérés comme non pertinents et rejetés. Cette étape permet de produire des termes candidats complexes, tout comme des termes candidats simples.

À la fin de cette étape, le syntagme maximal *oxygène transporté* est décomposé en sa tête *oxygène* et son modifieur *transporté* (figure 4.1c).

— **Etape 3** : Des mesures statistiques sont associées aux termes candidats simples ou complexes pour les ordonner et sélectionner les plus pertinents par rapport au domaine cible [Hamon et al., 2014]. Ces mesures incluent la fréquence et la C-Value [Maynard and Ananiadou, 2000].

4.2.2 Adaptation de YaTeA pour l'arabe

Pour adapter YaTeA aux textes de spécialité en MSA, nous avons d'abord suivi les pratiques traditionnelles en constitution de terminologie. Nous considérons donc qu'un terme doit contenir au moins un nom.

L'adaptation que nous avons mise en œuvre porte sur l'étape 1 (section 4.2.2) et l'étape 2 (section 4.2.3). La figure 4.2 résume le processus d'extraction de termes à partir de textes en MSA. Les textes ont été préalablement analysés morphologiquement et une étiquette morpho-syntaxique est associée à chaque mot. Ici, l'analyse morphologique et l'étiquetage morpho-syntaxique sont réalisés à l'aide de l'analyseur MADA+TOKAN [Roth et al., 2008].

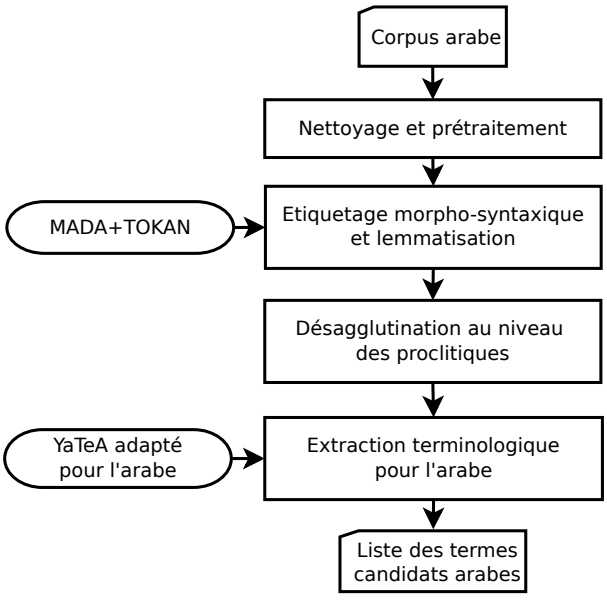


Fig. 4.2 – Processus d'extraction des termes arabes par adaptation de YaTeA

L'adaptation de l'étape 1 nous amène à définir les frontières syntaxiques. De manière similaire aux autres langues, les pronoms, les ponctuations et les verbes conjugués sont considérés comme des éléments ne pouvant pas apparaître dans les termes. Nous prenons également en compte certains éléments spécifiques de la langue arabe, tels que les pseudo-verbes (**كان**, **تكون**, **إن**) (*il était, elle soit, certes*), les adverbes relatifs [2] (**حيث**, **متى**, **كلما**) (*où* – avec une valeur relative de conjonction en arabe*, quand/au moment où/lorsque, au moment où/lorsque/chaque fois* – en même temps et en même proportion), ou les expressions lexicales ou les motifs, **في بعض الأوقات** (*parfois*), qui ne doivent pas non plus faire partie d'un terme.

L'arabe comporte certaines catégories syntaxiques, comme *le complément absolu* **المَفْعُولٌ المُطْلَقٌ**. Cependant, les mots de cette catégorie sont souvent étiquetés comme *Nom* par MADA+TOKAN. De ce fait, nous avons listé les lemmes de quelques compléments absolus dans la liste des frontières. Des mots comme *aussi* soulèvent des difficultés lorsqu'ils sont employés à côté d'un nom, comme dans ***Michaud aussi fut amoureux d'elle. Sainte-Beuve, Pensées et maximes,1869, p. 104.*** [3]. Bien qu'étant un adverbe, ce mot est associé ici, au moins en surface, au nom *Michaud*. Cela induit un étiquetage automatique autre qu'adverbe (ou encore : *... Alex, et Nathalie aussi*.

Ces frontières syntaxiques comportent 39 étiquettes morpho-syntaxiques (les noms propres : noun_prop, les pseudo-verbes : verb_pseudo, etc.) 28 mots sous leur forme fléchie (**ايضا** (*aussi*), **حسب** (*selon*), etc.) et 20 lemmes (**سَنَة** (*année*), **غَيْر** (*non/pas/autre que/sauf*), etc.), et 6 exceptions, c'est-à-dire des formes fléchies de frontières syntaxiques qui peuvent apparaître dans les termes candidats (**من** (*de/à partir de*), **في** (*en/dans*), etc.).

De plus, nous avons déterminé les étiquettes morpho-syntaxiques qui ne doivent pas figurer au début ou à la fin d'un terme. Il s'agit surtout de prépositions comme **من** (*de*), **إلى** (*à*), **بَيْنَ** (*entre*), **عند** (*lorsque*), auxquelles l'analyseur MADA+TOKAN attribue par erreur la catégorie morpho-syntaxique de nom. Enfin, nous avons défini des séquences de mots et d'étiquettes morpho-syntaxiques permettant d'éviter l'analyse de syntagmes nominaux maximaux mal formés par des patrons, lors de l'étape suivante comme un nombre suivi par *année* ou *an* **ثمان سنوات** (*huit ans*).

Ainsi, dans la phrase exemple présentée à la figure 4.3a, les frontières syntaxiques **تكون** (*elle soit*) (pseudo-verbe), **التي** (*que/laquelle*) (pronom relatif) , **يحملها** (*la transporte*) (verbe + complément d'objet direct) et **أقل** (*moins*) (adverbe) permettent d'identifier les syntagmes **كمية**

2. Catégorie syntaxique d'une sous-famille d'adverbes, étiquetés comme tels par MADA+TOKAN. En arabe, elle fait partie de la catégorie grammaticale **ظرف** *dharf* (*le conditionnel*).

3. `http://www.cnrtl.fr/definition/aussi`

الأكسجين (*quantité d'oxygène*) et الدم (*le sang*).

Pour l'adaptation de l'étape 2, nous avons défini les patrons réalisant l'analyse syntaxique en tête/modifieur des syntagmes nominaux maximaux identifiés précédemment. Il s'agit de séquences d'étiquettes morpho-syntaxiques et de prépositions permettant d'identifier le rôle syntaxique des composants des syntagmes nominaux. Ces patrons prennent en compte les caractéristiques morphologiques telles le genre, le nombre et le cas des constituants. En particulier, nous utilisons *al-'iḍāfah* qui marque l'état construit et le génitif pour l'analyse des syntagmes nominaux. De même les proclitiques sont pris en compte au sein des patrons, comme par exemple و (*et*) dans le patron `noun-m-s-a-c(Terme1)` و `noun-m-s-n-c(Terme2)`[4]. Nous avons défini 696 patrons permettant d'analyser des séquences de deux mots pleins, et 21 patrons pour les séquences de trois mots pleins. Ces patrons seront appliqués récursivement sur les syntagmes maximaux.

Par exemple, le patron `noun-m-s-g-d(Modifieur)` `noun-f-s-n-c(Tête)`[5] permet d'analyser le syntagme maximal كمية الأكسجين (quantité d'oxygène) tel que présenté à la figure 4.3c.

4.2.3 Prise en compte de phénomènes spécifiques à la langue arabe

4.2.3.1 Voyellation

Comme expliqué dans la section 1.3.2.1, les textes en langue arabe se caractérisent par la présence de symboles optionnels non alphabétiques appelés diacritiques, qui sont ajoutés au dessus ou au dessous d'une lettre. Un mot, et a fortiori un texte, peut donc apparaître sous deux formes : une forme non-voyellée فصل et une forme voyellée فَصَلَ (*il a licencié*) ou فَصْلٌ (*chapitre/section*). Ceci entraîne de nombreuses ambiguïtés et il est parfois difficile, voire impossible, de déduire le sens de certains mots non-voyellés si on ne connaît pas le contexte de leurs énonciations [Hadrich and Chaaben, 2006].

Comme la plupart des documents en langue arabe, notre corpus est constitué d'un ensemble de textes non diacritisés. L'absence de voyellation des textes du corpus provoque des erreurs d'étiquetage morpho-syntaxique et de lemmatisation dues aux ambiguïtés des formes non-voyellées, et en conséquence, une dégradation des résultats obtenus lors de l'extraction des termes. Pour cela, nous avons ajouté des traitements spécifiques dans le cas où deux mots

4. Le proclitique و est détaché du nom, car à ce stade on travaille sur un corpus désagglutiné.

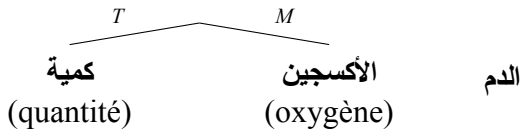
5. `noun-m-s-g-d` : nom masculin singulier défini au génitif. `noun-f-s-n-c` : nom féminin singulier construit au nominatif.

(6) تكون كمية الأكسجين التي يحملها الدم أقل
moins le-sang la-transporte que d'oxygène quantité pseudo-verbe
la quantité d'oxygène dans le sang est moindre

(a) Texte initial

(7) تكون كمية الأكسجين التي يحملها الدم أقل
O quantité d'oxygène O O le-sang O

(b) Étape 1 : Découpage à l'aide de frontières syntaxiques
(O représente un mot ne faisant pas partie d'un terme)



(c) Étape 2 : Analyse syntaxique des termes

Fig. 4.3 – Étapes d'extraction des termes candidats sur une phrase en arabe

s'écrivent de la même manière et possèdent la même catégorie morpho-syntaxique. Lors de la définition des frontières syntaxiques, nous avons utilisé les formes lemmatisées et voyellées. Aussi, nous avons dû prendre en compte ce phénomène dès la définition des patrons en utilisant des formes lemmatisées. MADA+TOKAN, que nous avons utilisé pour l'étiquetage morpho-syntaxique et la lemmatisation du corpus, associe un lemme sous sa forme voyellée à chaque mot du corpus.

4.2.3.2 Agglutination et clitiques

L'agglutination est une caractéristique élémentaire de la langue arabe. Elle consiste à intégrer des éléments particuliers du lexique, appelés clitiques, au mot auquel ils se rapportent. Nous avons introduit le phénomène d'agglutination, appelé aussi morphologie concaténative d'un mot arabe, dans la section 1.3.2.3.

Pour **معصميك** (*tes poignets* ou plus précisément *tes deux poignets*), on distingue le mot **معصمي** (dont la forme nominative est **معصمان** -- *deux poignets*) suivi par l'enclitique **ك** qui est un pronom possessif ((*ton, ta, tes*) -- dans cet exemple, il se traduit par ***tes***). De même, pour le mot **والأسنان**

(*et les dents*), nous distinguons les éléments suivants : **و** (*et*) un proclitique qui représente une conjonction de coordination, **ال** (*les*) un proclitique qui représente un article et le mot **أسنان** (*dents*).

Du point de vue du traitement automatique de l'arabe, il est parfois difficile de distinguer un proclitique ou enclitique d'un caractère du mot en question. Dans l'exemple suivant : **وسائل** deux analyses sont possibles ; i) on peut le considérer comme un seul mot : **وسائل** (*moyens*), et ii) comme il peut représenter un proclitique **و** (la conjonction de coordination ***et***) suivie du nom **سائل** (*liquide*).

Par conséquent, cette caractéristique engendre des ambiguïtés morphologiques lors de l'analyse. Ainsi, dans notre cas, ne pas traiter l'agglutination complique la création des patrons et provoque des erreurs lors de l'extraction des termes. Dans un premier temps, nous avons choisi de nous intéresser aux proclitiques et de les séparer des mots auxquels ils sont associés. Nous exploitons l'analyse morphologique réalisée par MADA+TOKAN. Celle-ci permet d'obtenir une décomposition des mots du corpus en séparant les proclitiques et les enclitiques. La prise en compte des proclitiques a nécessité la définition de 21 patrons d'analyse syntaxique supplémentaires comme par exemple :

`noun-f-s-n-d(Tête) (adj-f-s-g-i(Modifieur) noun-m-s-g-d(Tête))(Modifieur)`

ou `noun-m-s-a-c(Tête) و noun-m-s-n-c(Modifieur)`

4.2.3.3 Marques morphologiques du cas

En arabe, les noms peuvent avoir une des trois valeurs de cas selon la fonction syntaxique du mot dans la phrase. Autrement dit, à chaque fonction du mot correspond un cas qui sera marqué par la voyelle de la dernière lettre [Habash, 2010] :

— le nom au nominatif portera à la fin une *damma* : ُ ou *Dammatan* ٌ ;

— le nom à l'accusatif portera à la fin une *fatha* : َ ou *Fathatan* ً ;

— le nom au génitif portera à la fin une *kasra* : ِ ou *Kasratan* ٍ.

Les noms possèdent également trois états possibles [Habash, 2010] :

— l'état défini correspond à la forme nominale définie grâce à un article ;

— l'état indéfini désigne une instance non spécifique d'un nom ;

— l'état construit, quand deux noms sont liés ensemble pour former une relation de possession appelée en arabe *al-'iḍāfah* **الاضافة** (*possession*). Elle est formée d'une tête appelée

moudaf مُضافٌ (*possédé*) et d'un modifieur représentant un *moudaf ilayhi* مُضافَ إِلَيْهِ (*possesseur*) exprimant l'état construit toujours au cas génitif.

Dans le cadre de notre travail, les marques morphologiques des noms comme l'état et le cas sont des caractéristiques utiles pour identifier la fonction grammaticale du nom. Il est donc absolument nécessaire de prendre en considération ces caractéristiques pour analyser syntaxiquement les syntagmes extraits à partir des textes arabes. Ainsi, nous avons traité ces spécificités lors de la définition des patrons d'analyse syntaxique. Par exemple, pour déterminer une relation de possession (*al-'iḍāfah* الاضافة), la tête du patron doit être un nom à l'état construit, suivie par un nom au génitif. Dans ce contexte, le patron `noun-m-s-g-d(Modifieur) noun-f-s-n-c(Tête)` nous permet d'extraire le terme candidat كمية الأكسجين (quantité d'oxygène) tel que présenté à la figure 4.3c.

Actuellement, nous exploitons les étiquettes morpho-syntaxiques et les informations fournies lors de l'analyse morphologique de MADA+TOKAN. Nous évaluerons cette adaptation de YATEA dans la section 5.5.

4.3 Extraction de termes arabes par translittération

L'un des facteurs les plus importants qui ont contribué à la modernisation rapide de la langue arabe a été l'assimilation de mots spécialisés issus d'autres langues. Ainsi, à partir d'une étude réalisée sur notre corpus, nous constatons qu'il existe un nombre important de termes médicaux arabes qui sont créés par translittération de termes anglais.

Dans le cadre de la constitution d'une terminologie arabe dans le domaine médical, il nous est donc apparu important de prendre en compte ce phénomène. Pour cela, nous proposons une méthode spécifique permettant d'extraire de termes médicaux arabes issus de la translittération de termes anglais.

Dans un premier temps, nous devons constituer une liste des couples de termes candidats simples anglais-arabe. Les différentes étapes suivies pour la création de cette liste seront décrites à la section 4.3.1. Nous appliquons ensuite une méthode de détection des termes translittérés basée sur une table de correspondance des caractères anglais-arabe que nous avons définie. Nous détaillons cette méthode à la section 4.3.2. Finalement, l'identification des termes arabes nécessite l'enrichissement de notre processus d'extraction par des traitements complémentaires. Ceux-ci seront présentés à la section 4.3.3.

4.3.1 Construction de la liste des couples de termes anglais-arabe

Dans cette section, nous présentons l'ensemble des traitements nécessaires à la constitution de la liste des couples de termes simples anglais-arabe. Dans l'intention de détecter les termes candidats translittérés de l'anglais en caractères arabes, nous avons d'abord besoin d'une liste des termes anglais présents dans notre corpus. Pour cela, nous nous avons utilisé l'outil YATEA pour l'anglais afin d'extraire automatiquement des termes candidats de la partie anglaise du corpus.

Nous avons ensuite effectué un alignement au niveau des mots entre le corpus anglais et celui en arabe avec GIZA++. Le corpus arabe utilisé est constitué d'un ensemble de textes désagglutinés dont les enclitiques, les proclitiques et les articles sont séparés du mot auquel ils se rapportent (voir *Alignement3* de la section 3.3.2). Nous utilisons les résultats de l'alignement réalisé par GIZA++. Parmi les différents informations produites, nous utilisons la liste des correspondances des mots anglais avec ceux en arabe. Ainsi, pour chaque mot anglais, GIZA++ propose un ensemble de mots en arabe avec un taux de correspondance pour chaque couple. Ce fichier contient les tables inverses finales définies par un modèle des probabilités de traduction lexicale. Le taux de correspondance représente la probabilité que le mot M_s dans la langue source soit traduit en mot M_c dans la langue cible. La figure 4.4 présente un extrait de l'alignement fourni par GIZA++. Il s'agit des mots arabes que GIZA++ associe au mot anglais ***bleeding*** avec leurs taux de correspondance[6].

bleeding	0.904904	**نزيف**	(*saignement*)
bleeding	0.0543507	**بقعا**	(*tâches*)
bleeding	0.0407324	**اتصلي**	(*appelles*)
bleeding	1.2896e-05	**منتظم**	(*régulier*)

Fig. 4.4 – Liste des couples de correspondance pour le mot anglais ***bleeding*** (*saignement*)

Nous supposons que les alignements proposés par GIZA++ ayant les taux de correspondance les plus élevés comportent les termes arabes simples candidats parallèles à ceux en anglais. Aussi, à partir de la liste des correspondances fournie par GIZA++, nous avons extrait, pour chaque mot en anglais, le couple de mots anglais-arabe ayant le taux de correspondance le plus élevé. Par exemple, pour le mot anglais ***bleeding*** nous avons gardé le mot arabe **نزيف** (*saignement*). La figure 4.5 présente un extrait de la liste des couples des mots ayant le taux de correspondance le plus élevé[7].

6. Nous avons ajouté les traductions des mots arabes en français pour une meilleure compréhension par le lecteur.

7. Nous avons ajouté les traductions des mots arabes en français pour aider la lecture.

male	ذكور	(*mâles*)
knees	ركبتان	(*deux genoux*)
ligament	رباط	(*ligament*)
conditions	احوال	(*conditions*)
reflex	منعكس	(*reflété*)
aches	احساس	(*déchets*)
drug	صيدلية	(*pharmacie*)

Fig. 4.5 – Liste des couples de mots alignés anglais-arabe ayant le taux de correspondance le plus élevé

Il est important de mentionner que la liste des couples des mots anglais-arabe extraits ayant le taux de correspondance le plus élevé ne présente pas toujours des alignements corrects ou ceux les plus exacts. Par exemple, pour le mot anglais ***aches*** (*douleurs*), notre méthode lui a associe le mot arabe احساس (*sentiment*) dont le taux de correspondance est 0,999314 plutôt que اوجاع (*douleurs*) qui a un taux de 0,200076. Ceci présente une des erreurs d'alignement produites par GIZA++. Finalement, nous repérons les termes anglais présents dans la liste de correspondance et nous constituons notre liste de couples de termes simples anglais-arabe. L'évaluation de cette liste sera présentée à la section 5.6 du chapitre 5.

4.3.2 Méthode proposée

Après avoir créé la liste des couples de termes candidats simples anglais-arabe, nous parcourons cette liste pour en extraire les termes candidats arabes produits par translittération de termes anglais. Dans cette section, nous décrivons les différentes étapes constituant la méthode que nous proposons. La figure 4.6 résume le processus mis au point.

Lemmatisation des noms Même si les termes arabes transcrits proviennent d'autres langues étrangères, ceux-ci suivent des règles flexionnelles imposées par la langue arabe pour la production des formes du pluriel des noms et des adjectifs. Cependant, pour obtenir des résultats de meilleure qualité, il est souhaitable de disposer de la forme au singulier des termes anglais et arabes. Pour les termes candidats anglais, nous avons utilisé les lemmes associés aux mots par TreeTagger. Pour les termes candidats arabes, nous avons dû définir des règles de lemmatisation décrites ci-dessous étant donné que MADAMIRA ne propose pas de lemmes pour la plupart des mots empruntés à une autre langue.

Dans la littérature arabe, le passage du singulier au pluriel pour les noms empruntés est réalisé en ajoutant la terminaison ات (*ate/at/èt/ete*). Nous testons donc si les termes arabe et anglais,

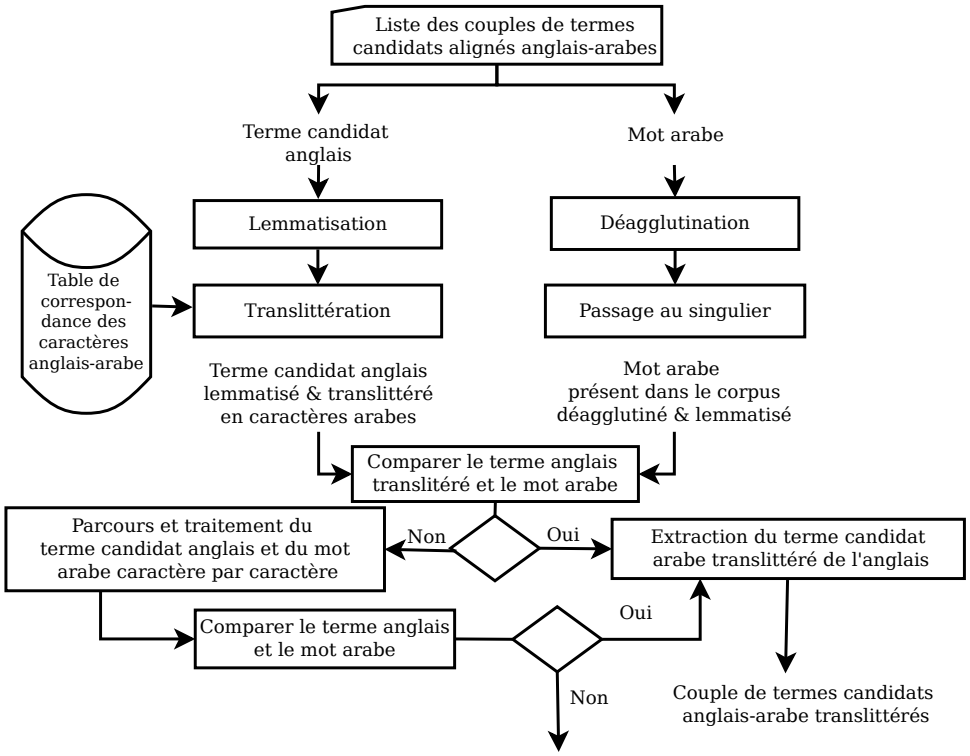


Fig. 4.6 – Processus d'extraction des termes arabes translittérés

n'ayant pas de lemmes offerts respectivement par MADAMIRA ou TreeTagger, sont marqués par **ات** et ***s*** respectivement. Dans ce cas, nous supprimons les signes du pluriel et nous gardons le reste comme étant la forme au singulier du terme. En revanche, ceci ne peut pas présenter une règle générale car la terminaison ***s*** à la fin d'un terme ne marque pas toujours sa forme au pluriel. Par exemple, le terme ***arthritis*** (*arthrite*) présente la forme au singulier du terme. où cette terminaison fait partie du lemme. Dans le cas où seul le terme arabe a la terminaison du pluriel, nous vérifions alors si cette terminaison ne convient pas à une translittération des derniers caractères du terme en anglais.

Par exemple, prenons le couple de termes ***etomidate*** - **إيتومِيدات** ***étomidate***[8]. Nous constatons la présence de la terminaison du pluriel **ات** dans **إيتومِيدات** *itoumidat* à la fin du terme arabe. Or, en vérifiant la terminaison du terme anglais, nous apercevons qu'il s'agit bien du résultat de la translittération du terme anglais.

La lemmatisation des noms nous permet ainsi d'identifier le couple de terme ***protein*** - **بروتين** *(protéine)* dont le terme arabe apparaît dans le corpus sous la forme au pluriel **بروتينات** - *proutiy-net (des protéines)*. De même, pour le couple ***hormone*** - **هورمون** *(hormone)*, chacun des termes apparaît dans le corpus sous la forme du pluriel : *hormones* pour la partie anglaise et **هرمونات** -

8. L'étomidate est un médicament.

hormouunet pour celle en arabe *(des hormones)*.

Traitement de l'agglutination L'agglutination est une caractéristique essentielle de la langue arabe qui désigne la complexité de la structuration des mots arabes par leur morphologie concaténative (section 1.3.2.3). Comme nous l'avons présenté à la section 3.3, notre travail repose sur un alignement au niveau des mots effectué entre notre corpus anglais et celui en arabe désagglutiné au niveau des enclitiques, des proclitiques et des articles. Cependant, lorsqu'il s'agit d'un terme emprunté à une langue étrangère, l'étiqueteur morpho-syntaxique MADAMIRA ne parvient pas à identifier ou à séparer les articles et les proclitiques du mot en question. Leur présence empêche alors la détection de ces termes produits par translittération.

Prenons, par exemple, le terme médical anglais ***titanuim*** dont le terme correspond en arabe est **تيتانيوم** *titanioum* (*(titanium)*). À partir de notre liste initiale de couples de termes candidats, nous repérons le couple anglais-arabe ***titanium*** - **التيتانيوم** *altitanioum* (*(titanium - le titanium)*). Notre première étape est de translittérer la partie anglaise en se basant sur notre table de correspondance des caractères anglais en arabe. L'exemple (8) présente le terme candidat anglais ainsi que sa translittération obtenue grâce à notre table de correspondance des caractères anglais-arabe.

تيتانيوم titanuim (8)

Terme candidat anglais m u i n a t i t

Translittération en caractères arabes ت ي ت ا ن ي و م

Par la suite, nous avons effectué une comparaison du terme candidat arabe présent dans la liste de couples de termes et la translittération du terme anglais correspondant. Cette comparaison des deux termes **التيتانيوم** (*le titanium*) et **تيتانيوم** (*titanuim*) conduit à un rejet. Ainsi, nous avons remarqué que la présence des clitiques provoque des erreurs lors de l'extraction de termes anglais translittérés en arabe. La présence de l'article **ال** - *al* (*le*) dans **التيتانيوم** - *altitanioum* (*le titanium*) fait obstacle à notre méthode.

Sans procéder à une étape supplémentaire de désagglutination, la présence des termes arabes translittérés dans leurs formes définies entraîne leur élimination lors de la phase de comparaison des mots (voir figure 4.6). La prise en compte de la présence éventuelle des articles et des proclitiques a nécessité l'ajout de traitements supplémentaires au sein de notre processus d'extraction. Il s'agit d'identifier les articles et les prépositions situés au début des termes arabes pour avoir

une meilleure translittération du terme anglais en caractères arabe. Dans la langue arabe, l'intégration de ces éléments au mot auquel ils se rapportent se fait dans un ordre bien précis (voir section 1.3.2.3). Nous avons utilisé cet ordre pour définir des règles de désagglutination pour ces cas de figure.

Ainsi, pour chaque couple de termes, nous cherchons s'il existe des morphèmes au début du terme arabe qui peuvent représenter des proclitiques comme conjonction **و** (*et*), préposition **ل**, **ب** (*pour, par*), article **ال** (*le, la, les*) , etc. suivant cet ordre. Par la suite, nous vérifions la présence de la translittération de ces éléments au début du terme en anglais. S'il existe une correspondance entre les deux termes alignés, nous considérons alors ces éléments comme partie du terme. Sinon, nous les considérons comme proclitiques et nous les éliminons du terme arabe en question.

Par exemple, dans le couple ***titanium*** - **التيتانيوم**, nous remarquons l'existence de **ال** *al* au début du terme arabe. Or, le terme anglais ne commence pas par *'al'*. Nous considérons donc ce morphème comme un proclitique désignant un article défini et nous l'éliminons du terme arabe. Les exemples (9) et (10) montrent les ressemblances et les différences entre le terme candidat anglais et le terme correspondant en arabe, avec et sans désagglutination.

(9) titanuim التيتانيوم

Terme candidat anglais m u i n a t i t - -

Translittération de son parallèle agglutiné m u i n a t i t l a

Terme candidat arabe agglutiné ا ل ت ي ت ا ن ي و م

(10) titanuim تيتانيوم

Terme candidat anglais m u i n a t i t

Translittération de son parallèle désagglutiné m u i n a t i t

Terme candidat arabe désagglutiné ت ي ت ا ن ي و م

Il existe toutefois des exceptions. Par exemple, dans **الومينيوم** *alouminioum* ((*aluminium*)), même si ce terme arabe commence par **ال** *al*, nous considérons cet élément comme faisant partie du terme car le terme anglais correspondant ***aluminium*** commence aussi par *al*. L'exemple (11) présente la correspondance entre ces deux termes.

(11) aluminium الومينيوم

Terme anglais m u i n i m u l a

Terme arabe translittéré m u i n i m u l a

Terme arabe ا ل و م ي ن ي و م

Le tableau 4.1 présente quelques exemples de termes arabe extraits après l'étape de désagglutination décrite ci-dessus.

Couple de termes proposé par GIZA++		Terme arabe translittéré
Terme anglais	Terme arabe aggluti né	
titanium	**التيتانيوم** (*le titanium*)	**تيتانيوم** *(titanium)*
progesterone	**والبروجسترون** *(et le progestérone)*	**بروجسترون** *(progestérone)*
steroid	**بالاستيرويد** *(par le stéroïde)*	**ستيرويد** *(stéroïde)*

Tab. 4.1 – Exemple de termes arabes extraits suite au traitement de l'agglutination

Translittération du terme anglais en caractère arabe Pour extraire ces termes arabes, nous avons créé notre propre table de correspondance des caractères anglais vers l'arabe. Pour cela, nous nous sommes appuyés sur une étude de notre corpus médical parallèle anglais-arabe afin de décrire les mécanismes de translittération d'un mot anglais vers un mot arabe. Nous nous sommes également inspirés des règles de translittération des caractères arabes en caractères latins fixées par la norme ISO 233-2 (1993). Notre objectif est d'obtenir une table de correspondance adaptée aux caractéristiques phonologiques utilisées lors du passage de l'anglais à la langue arabe. Notre corpus arabe n'étant pas voyellé, nous avons choisi de ne pas prendre en considération la correspondance des voyelles courtes arabes.

La table 4.2 présente un extrait de notre table de correspondance. Pour chaque caractère anglais, nous rassemblons les différents caractères arabes qui peuvent lui correspondre. Les caractères arabes y sont ordonnés en fonction de la correspondance la plus fréquente avec le caractère anglais auquel ils se réfèrent. Nous avons défini cet ordre à la suite de l'étude préliminaire réalisée sur le corpus.

Par exemple, le caractère ***g*** peut être translitteré par le caractère **ج** comme dans le couple ***g****lucophage* - **جلوكوفاج** ou par le caractère **غ** comme dans le couple ***g****luconate* - **غلوكونات**.

Caractère anglais	Caractères arabes
g	غ ج
r	ر
c	ق ك س

Tab. 4.2 – Extrait de la table de correspondance des caractères

Le tableau 4.3 présente quelques exemples des couples de termes extraits grâce à la table de correspondance que nous avons définie.

Terme anglais	Terme arabe	Transcription	Synonyme français
glucophage	**جلوكوفاج**	*jloukoufaj*	(*glucophage*)
vitamin	**فيتامين**	*vitamin*	(*vitamine*)
cream	**كريم**	*criym*	(*crème*)
tetanus	**تيتانوس**	*titanous*	(*tétanos*)
barium	**باريوم**	*barioum*	(*baryum*)
typhus	**تيفوس**	*tifous*	(*typhus*)

Tab. 4.3 – Exemples de couples de termes obtenus à partir de notre table de correspondance des caractères

Certains termes arabes peuvent être directement extraits grâce à notre table de correspondance lorsque les caractères du terme anglais correspondent à un seul caractère en arabe ou lorsqu'ils sont translittérés par le premier caractère arabe équivalent. Cependant, l'identification de la plupart des termes arabes nécessite l'enrichissement de notre processus d'extraction par des traitements complémentaires. Nous décrivons ces traitements dans la section suivante.

4.3.3 Traitements complémentaires

Plusieurs termes arabes produits par translittération de l'anglais en caractères arabes nécessitent des traitements complémentaires pour assurer leur détection. Ces traitements tiennent compte des particularités de la langue arabe provenant de la conversion des voyelles de l'anglais. Comme la plupart des documents en langue arabe, notre corpus est constitué d'un ensemble des textes non diacritisés. Cette non-voyellation provoque l'apparition de différentes formes de translittération d'un même terme anglais. Par conséquent, certains termes anglais peuvent correspondre à plusieurs termes arabes qui diffèrent sur la forme mais qui se prononcent de la même manière.

Ainsi, lors du passage des termes anglais à la langue arabe par translittération, certains terminologues tentent de remplacer les voyelles latines par des voyelles arabes courtes, alors que d'autres mettent l'accent sur ces voyelles en les remplaçant par des voyelles arabes longues.

Les signes diacritiques des voyelles courtes sont alors suivis par les lettre de prolongation pour désigner les voyelles longues qui sont *alif* ا, *waw* و et *ya* ي (voir 1.3.2.1).

Par exemple, la translittération du terme ***bacteria*** (*bactérie*) est présentée en arabe soit par la forme **بكتريا** - *bktria* soit par **بكتيريا** - *bktiria*. Dans le premier terme arabe, la voyelle ***e*** de ***bacteria*** est remplacée par la voyelle arabe longue 'ي'[9] Les exemples 12 et 13 montrent les deux translittérations possibles du mot anglais ***bacteria*** en caractères arabes.

bacteria (12)

a i r te c ba
ب ك ت ر ي ا
بكتريا

bacteria (13)

a i r e t c ba
ب ك ت ي ر ي ا
بكتيريا

De même, pour le terme ***oxygen*** - (*oxygène*) dont la transcription en arabe apparaît trois fois dans notre corpus sous la forme **اوكسجين** - *ouksjin* où la voyelle ***o*** est représentée par le caractère ا suivi par la voyelle longue و et six fois sous la forme **اكسجين** - *oksjin* où la voyelle longue و n'est pas présente. Ceci montre qu'un terme anglais translitteré peut être écrit de plusieurs manières différentes en caractères arabes. Les exemples (14) et (15) montrent les deux translittérations possibles du mot anglais ***oxygen*** en caractères arabes.

oxygen (14)

n ge xy o
او كس جي ن
اوكسجين

9. La différence graphique est liée au changement des allographes des caractères arabes selon leur position dans un mot ou leur position indépendante (lorsque le caractère est représenté seul).

oxygen (15)

n ge xy o

ا كس جي ن

اكسجين

Pour chaque couple de termes proposé par GIZA++, nous avons testé, dans un premier temps, la similitude du terme anglais translittéré à partir de notre table de correspondance des caractères et le terme arabe tel qu'il est présent dans le corpus après son passage à la forme au singulier et l'étape du traitement de l'agglutination. Si la comparaison des deux mots a conduit à un rejet, nous procédons aux traitements complémentaires en analysant ces deux mots caractère par caractère. Voici les règles à suivre :

— Si les deux caractères arabes sont similaires alors passer au caractère suivant ;
Sinon,

— si le caractère anglais correspond aux deux caractères arabes alors remplacer le caractère du terme anglais translittéré par celui du terme arabe issu dans le corpus ;
Sinon,

— si l'un des caractères arabe correspond à une voyelle arabe longue **et**

  — le caractère suivant correspond au caractère en question du deuxième mot ;
  ou

  — le caractère suivant correspond au même caractère anglais que celui du deuxième mot ;

  alors suivre le modèle du terme arabe ;

Prenons, par exemple, le couple de termes suivant ***gluconate*** - **الغلوكونات** (*(gluconate)*). Après avoir appliqué les traitements complémentaires de l'agglutination, nous obtenons le couple ***gluconate*** - **غلوكونات** en supprimant l'article défini **ال** situé au début du terme arabe (comme nous l'avons détaillé dans la section 4.3.2). L'exemple (16) montre le résultat de la translittération du terme anglais à partir de notre table de correspondance des caractères et avant l'application des traitements complémentaires.

gluconate (16)

e t a n o c u l g

ج ل و ك و ن ا ت

جلوكونات

Nous comparons alors si le terme arabe tel qu'il existe dans le corpus **غلوكونات** (désagglutiné) est celui que nous avons obtenu par la translittération **جلوكونات**. Clairement, il ne s'agit pas du même mot arabe puisqu'ils diffèrent au niveau du premier caractère. Nous réalisons alors les traitements complémentaires. L'exemple (17) présente une comparaison entre les caractères du terme arabe et celui translitteré.

غلوكونات (17)

غ ل و ك و ن ا ت Terme arabe

ج ل و ك و ن ا ت Terme translitteré

جلوكونات

Dès la première itération, nous remarquons que les premiers caractères ne sont pas identiques. Or, comme le montre le tableau 4.2, le caractère anglais ***g*** correspond aux deux caractères arabes **ج** et **غ**. Dans ce cas, le caractère arabe **ج** sera remplacé par le caractère **غ** pour respecter la forme du terme tel qu'il existe dans notre corpus.

Soit un autre exemple, le couple de termes ***hormone*** - **هرمون**. Nous avons bien vérifié que le terme arabe n'est pas au pluriel et ne contient pas de proclitique au début du mot. Notre translittération du terme anglais produit le terme translitteré **هورمون**. L'exemple (18) présente une comparaison entre les caractères du terme arabe et ceux du terme translitteré.

هرمون (18)

ه - ر م و ن Terme candidat arabe

ه و ر م و ن Terme anglais translitteré

هورمون

Comme le montre l'exemple, les caractères en seconde position sont différents. Nous remarquons que le deuxième caractère du terme translittéré correspond à une voyelle longue و et celui d'après (le troisième caractère du terme translittéré) est identique au deuxième caractère du terme arabe. Dans ce cas, nous supprimons la voyelle longue و située à la deuxième position du terme translittéré. La comparaison finale de ces deux termes arabes après application des traitements complémentaires montre qu'il s'agit bien d'un terme anglais translittéré en caractères arabes. Nous présentons et discutons les résultats obtenus dans le chapitre 5 (section 5.6).

4.4 Extraction des termes candidats arabes par transfert

Adapter une méthode existante ou mettre au point une méthode spécifique à une langue pour extraire automatiquement des termes demande un effort important d'analyse et de conception. Afin de limiter ce coût, nous avons également choisi de proposer une méthode d'acquisition terminologique pour la langue arabe se basant sur la notion de transfert translingue [McDonald et al., 2011]. Il s'agit de mettre en œuvre un processus d'extraction de termes à partir des textes d'une langue source (ici l'anglais) puis de transférer les informations extraites sur des textes d'une langue cible (ici, l'arabe standard moderne) afin d'identifier le même type d'informations terminologiques. Ainsi, à partir d'un corpus de spécialité parallèle anglais-arabe, les termes arabes sont extraits grâce aux relations de traduction pouvant exister entre les deux langues. Nous supposons alors que les expressions arabes qui correspondent aux termes candidats anglais extraits automatiquement, représentent des termes candidats arabes.

Outre la réduction du coût de mise au point d'une méthode d'extraction de termes spécifiques à une langue, cette approche par transfert permet également de palier l'absence de ressources ou d'outils de TAL dans une langue donnée. La grande majorité des traitements est réalisée sur le corpus source et s'appuie sur un alignement au niveau des mots des textes sources et cibles (section 4.4.1). Les interventions sur le corpus cible se limitent à une projection des termes candidats anglais par l'intermédiaire de l'alignement réalisé au niveau des mots sur les couples de phrases parallèles (section 4.4.2). La figure 4.7 présente notre méthode d'extraction terminologique proposée basée sur la notion de transfert translingue.

4.4.1 Traitements préliminaires

Dans cette section, nous présentons les différentes étapes et traitements nécessaires avant de procéder à la projection des termes anglais sur les textes arabes grâce aux alignements au

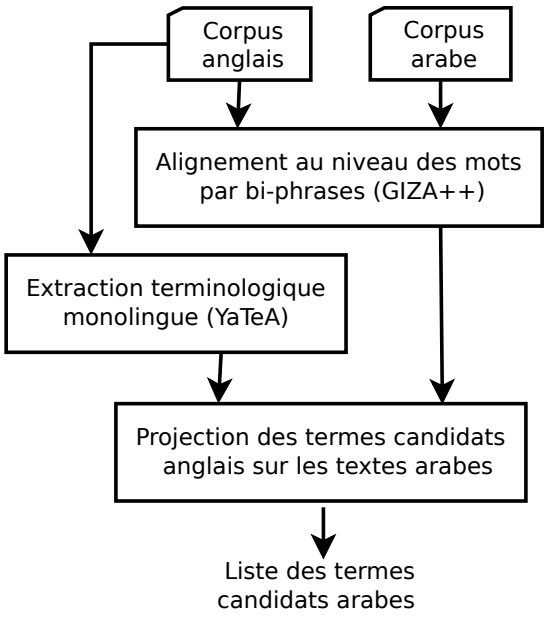


Fig. 4.7 – Processus d'extraction des termes arabes par transfert translingue

niveau des mots réalisés sur notre corpus anglais-arabe.

Extraction terminologique à partir du corpus source Le corpus source, ici en anglais, subit les traitements habituels permettant d'extraire des termes candidats. Après un étiquetage morpho-syntaxique et une lemmatisation réalisés par TreeTagger Schmid [1997], nous effectuons une extraction des termes candidats anglais. Pour cela, nous avons utilisé l'outil YaTeA (voir section 4.2.1).

Alignement du corpus parallèle Une étape d'alignement est indispensable à la mise en place d'une méthode de transfert. Il s'agit d'aligner notre corpus parallèle anglais-arabe au niveau des mots en suivant les différentes étapes d'alignement détaillées à la section 3.3 du chapitre 3. Nous utilisons ainsi notre corpus arabe désagglutiné tel que réalisé par la méthode *Alignement3* de la section 3.3.2 : les enclitiques, les proclitiques et les articles sont séparés du mot auquel ils se rapportent.

Parmi les différents résultats produits par GIZA++, nous avons choisi d'utiliser les fichiers d'extension `.A3.final`, car ceux-ci proposent le plus d'informations utiles pour l'alignement des couples de phrases parallèles anglais-arabe. Le format du fichier d'alignement est illustré dans la figure 4.8.

Pour chaque paire de phrases bilingues, l'alignement est représenté par trois lignes : la première ligne contient des informations générales sur ce couple comme le nombre séquentiel de

```
# Sentence pair (1) source length 3 target length 3 alignment score :
0.00116445
about your pain
NULL ({ }) ({ 2 }) ك ({ 3 }) الم ({ 1 }) عن
```

Fig. 4.8 – Extrait du résultat de l'alignement, utilisé pour le transfert

phrases dans le corpus, la longueur des phrases et la probabilité d'alignement ; la deuxième ligne est la phrase en anglais telle qu'elle figure dans le corpus ; la troisième ligne comporte la phrase en arabe en attribuant à chaque mot un ensemble de valeurs entières correspondant à la position des mots anglais avec lesquels le mot arabe est aligné.

4.4.2 Projection des termes candidats anglais sur les textes arabes

Après avoir extrait les termes candidats de la langue anglaise et obtenu un alignement au niveau des mots de notre corpus parallèle anglais-arabe par couple de phrases, nous projetons ces termes sur le corpus arabe. Pour cela, nous utilisons la liste des termes candidats extraits par YaTeA. Nous tenons à mentionner que cette liste n'indique pas la position où ces termes ont été trouvés dans le corpus. Notre objectif est d'extraire les termes candidats arabes sans avoir recours à un extracteur de termes pour cette langue.

Tout d'abord, nous effectuons un parcours du résultat de l'alignement par couple de phrases parallèles. À partir de la liste des termes candidats anglais, simples et complexes, nous vérifions si certains d'entre eux figurent dans la phrase en anglais. Nous récupérons ainsi leurs positions dans la phrase. Ensuite, nous cherchons si ces positions correspondent à un ou plusieurs mots en arabe grâce à l'alignement proposé par GIZA++ et décrit dans la section 4.4.1. Nous identifions alors tous ces mots dans la phrase parallèle en arabe et nous délimitons la partie textuelle minimale qui les comporte. Nous identifions, dans un premier temps, les frontières qui correspondent aux mots à leurs extrémités. Ensuite, nous retirons l'ensemble de mots délimité par ces deux extrémités.

Nous supposons que T_s le terme médical anglais extrait par YaTeA est composé de l'ensemble des mots $T_s = \{ m_{s1} , \ldots , m_{sn} \}$ et $Corr_{Ar} = \{ m_{ar1} , \ldots , m_{arp} \}$ l'ensemble des mots constituant la projection de T_s (c'est-à-dire l'union des projections de chaque mot dans T_s). Nous nous considérons que la projection du terme médical anglais T_s est la séquence ordonnée de mots $T_c = \{ m_{c1} , \ldots , m_{ck} \}$ dans la phrase arabe telle que :

— $m_{c1} \in Corr_{Ar}$ et $position(m_{c1}) = positionMin(Corr_{Ar})$[10]

— $m_{ck} \in Corr_{Ar}$ et $position(m_{ck}) = positionMax(Corr_{Ar})$[11]

Dans ce qui suit, nous présentons sur un exemple, les différentes étapes de notre méthode d'extraction terminologique basée sur la notion de transfert translingue. Pour cela, nous avons choisi la phrase anglaise ainsi que son équivalent en arabe sous sa forme désagglutinée figurant à la figure 4.9.

Version anglaise	dental x-rays are safe if a shield is placed over your abdomen .
Version arabe	**والأشعة السينية على الأسنان آمنة إذا تم وضع حجاب واقي على البطن**
Version arabe (désagglutinée)	**و ال اشعة السينية علي ال اسنان امنة اذا تم وضع حجاب واقي علي ال بطن**
Version française	Les radiographies dentaires sont sures si un écran est disposé sur votre abdomen.

Fig. 4.9 – Phrase en anglais et les phrases correspondantes en arabe, désagglutinées ou non, et en français

L'exemple 19 présente la traduction en français des mots et des clitiques qui constituent la phrase arabe désagglutinée.

(19) والأشعة السينية على الأسنان آمنة إذا تم وضع حجاب واقي على البطن

و ال اشعة السينية علي ال اسنان امنة اذا تم وضع حجاب واقي

protecteur voile placement effectué si sans-danger dents les sur X rayons les et

علي ال بطن

abdomen l' sur

et les rayons X sur les dents sont sans danger si un voile de protection est placé sur l'abdomen

Après avoir réalisé l'étape d'alignement, nous obtenons les correspondances produites pour chaque couple de phrases parallèles entre les mots qui les constituent. La figure 4.10 présente un extrait du fichier résultat produit par GIZA++ pour notre exemple (figure 4.9).

Afin d'extraire les termes candidats arabes parallèles, nous effectuons une projection de ces termes anglais sur la partie arabe grâce à l'alignement obtenu précédemment. La figure 4.11 présente les termes candidats anglais extraits dans notre exemple précédent ainsi que leurs positions dans la phrase.

10. la plus petite position en nombre de mot arabe dans la phrase arabe parallèle
11. la plus grande position en nombre de mot arabe dans la phrase arabe parallèle

```
# Sentence pair (198) source length 16 target length 13 alignment score :
3.61294e-19
dental x-rays are safe if a shield is placed over your abdomen .
NULL ({ 12 }) ({ 11 6 }) بطن ({ }) ال ({ }) على ({ 13 }) واقي ({ 10 9 8 7 }) حجاب
({ }) وضع ({ }) تم ({ 5 }) إذا ({ 4 }) آمنة ({ 1 }) اسنان ({ }) ال ({ }) علي ({ 3 })
السينية ({ 2 }) اشعة ({ }) ال ({ }) و
```

Fig. 4.10 – Résultat de l'alignement

dental	1	(*dentaire*)
x-rays	2	(*rayons X*)
shield	7	(*écran*)
dental x-rays	1 2	(*rayons X dentaires*)
abdomen	12	(*abdomen*)
safe	4	(*sans danger*)

Fig. 4.11 – Liste des termes candidats anglais extraits et leurs positions

Lorsqu'il s'agit des termes candidats anglais simples, nous récupérons les mots arabes auxquels les positions correspondent et nous les considérons en tant que leurs termes arabes candidats simples. La figure 4.12 présente les termes candidats anglais simples et leurs correspondant en arabe.

dental	**اسنان**	(*dents*)
x-rays	**اشعة**	(*rayons*)
shield	**حجاب**	(*voile*)
abdomen	**بطن**	(*abdomen*)
safe	**امنة**	(*sans-danger*)

Fig. 4.12 – Liste des termes candidats anglais simples extraits
et leurs correspondants en arabe

Pour les termes complexes candidats anglais, nous récupérons d'abord les positions des mots qui les constituent comme le montre la figure 4.11. Puis, l'ensemble des mots arabes sont extraits grâce à la projections des mots anglais dans la phrase arabe grâce aux alignements proposés par GIZA++ : nous récupérons les mots de la phrase bornée par la position minimale et maximale des mots du terme anglais correspondant. Ainsi, pour le terme complexe ***dental x-rays***, nous récupérons les positions des mots qui le constituent : ***dental*** à la position *1* et ***x-rays*** à la position *2*. Par la suite, nous repérons les mots arabes qui leur correspondent : **اسنان** et **اشعة**. Nous récupérons alors les mots constituant le segment lexical minimal qui englobe ces mots.

Nous extrayons tous les mots à partir de **اشعة** (ayant la position minimale) jusqu'à **اسنان** (ayant la position maximale). Nous obtenons alors **اشعة السينية علي ال اسنان** (*littérale : rayons X sur les dents*) comme terme candidat arabe correspondant au terme candidat anglais en question. Nous ajoutons des traitements complémentaires par la suite afin de rattacher les articles et les prépositions au mot auquel il se rapportent. Nous obtenons finalement le couple de terme anglais-arabe ***dental x-rays*** **- اشعة السينية علي الاسنان** (*les radiographies dentaires*).

4.5 Conclusion

Dans ce chapitre, nous avons présenté trois stratégies permettant d'extraire des termes médicaux en arabe standard moderne (MSA). Nous nous sommes servis pour cela de notre corpus parallèle anglais-arabe. Dans un premier temps, nous avons proposé une méthode permettant d'adapter l'extracteur YATEA à l'arabe tout en prenant en considération différentes caractéristiques spécifiques à cette langue comme l'agglutination et la non voyellation. Par la suite, nous avons profité du fait que la langue arabe a adopté dans son lexique des termes d'une autre langue. Pour cela, nous avons construit un système qui nous permet de détecter et d'extraire les termes médicaux arabes empruntés à la langue anglaise. Puisqu'à notre connaissance, il n'existe pas de table de correspondance des caractères latins en arabe, nous avons construit notre table de correspondance des caractères anglais vers les caractères arabes. Finalement, nous avons proposé une méthode d'extraction terminologique pour l'arabe basée sur la notion de transfert translingue. Celle-ci exige d'avoir à notre disposition un corpus parallèle anglais-arabe aligné au niveau des mots ainsi qu'une liste de termes médicaux candidats extraite à partir des textes en anglais.

La qualité des travaux présentés ci-dessus va être évaluée dans le chapitre suivant à travers plusieurs expériences et une évaluation s'appuyant sur un protocole d'évaluation que nous avons proposé.

Chapitre 5

Évaluation : résultats et discussion

Sommaire

5.1 Introduction

Ce chapitre présente les résultats des différentes expériences qui mettent en œuvre les méthodes d'acquisition terminologique pour l'arabe standard moderne (MSA) décrites au chapitre 4. Pour être exhaustif sur l'évaluation des méthodes proposées, nous évaluons également la qualité des différents types d'alignements proposés au chapitre 3 ainsi que la qualité des termes extraits par YATEA sur le corpus anglais. Pour mener à bien l'analyse des résultats obtenus, nous utilisons un protocole d'évaluation général qui est ensuite adapté pour tenir compte des objectifs de chaque méthode d'acquisition terminologique proposée.

La section 5.2 présente l'évaluation des alignements proposés pour notre corpus parallèle anglais-arabe. Ensuite, nous proposons les principes généraux de notre protocole d'évaluation (section 5.3). Avant de décrire les expériences réalisées en vue de construire d'extraire des termes médicaux arabes, nous présentons à la section 5.4 une évaluation de l'outil YATEA sur notre corpus anglais. Cette évaluation préliminaire nous permettra de mieux cerner les sources d'erreurs potentielles, lorsque les termes candidats anglais sont utilisés pour extraire des termes arabes. Nous évaluons ensuite les résultats d'extraction terminologique produits par les différentes méthodes proposées : l'adaptation d'un extracteur de termes pour le MSA (section 5.5), l'extraction des termes anglais translittérés en caractères arabes (section 5.6) et l'extraction terminologique pour le MSA basée sur la notion de transfert translingue (section 5.7).

5.2 Alignement du corpus au niveau des mots

Les méthodes décrites dans le chapitre 4 s'appuyant sur un alignement au niveau des mots, nous avons voulu identifier l'analyse morpho-syntaxique et les pré-traitements de notre corpus qui permettent d'obtenir un alignement de bonne qualité. Pour cela, nous nous sommes concentré sur les phénomènes d'agglutination et l'impact de leur prise en compte sur l'alignement proposé par GIZA++. Ainsi, comme décrit dans la section 3.3.2, nous avons réalisé trois expériences :

— **Alignement1** : corpus arabe où les mots sont agglutinés.

— **Alignement2** : corpus arabe avec une désagglutination partielle des mots. Les textes sont désagglutinés au niveau des enclitiques et des proclitiques sauf les articles.

— **Alignement3** : corpus arabe avec une désagglutination totale des mots. Les textes sont désagglutinés au niveau des enclitiques et des proclitiques y compris les articles.

Dans ces trois expériences, pour chaque mot anglais, un mot arabe lui est associé avec le taux de correspondance le plus élevé proposé par GIZA++. Notre évaluation est basée sur ces probabilités de correspondance. Les trois expériences ont été effectuées sur 3917 mots anglais. Le premier alignement avec le corpus arabe agglutiné a produit 27800 couples de mots anglais-arabe contenant 9669 mots arabes. Comme le montre le tableau 5.1, nous avons remarqué qu'à chaque étape de la désagglutination, le nombre des alignements des mots diminue ainsi que le nombre des mots en arabes. Ceci s'explique par la réduction de la variété de formes des mots arabes lorsque les mots sont désagglutinés.

	Alignement1	Alignement2	Alignement3
Nombre de couples de mots	27800	20235	18190
Nombre de mots anglais	3917	3917	3917
Nombre de mots arabes	9669	6435	5395

Tab. 5.1 – Résultats des trois types d'alignements avant sélection des meilleures propositions

Chaque mot anglais étant aligné avec au moins un mot arabe, nous avons choisi de sélectionner les alignements possédant la probabilité de correspondance la plus élevée. A partir de cette sélection, nous avons calculé la moyenne des taux de correspondance pour chaque expérience (cf. tableau 5.2). Nous observons une amélioration des taux de correspondance au fur et à mesure de la prise en compte de l'agglutination au sein du corpus arabe. Nous considérons qu'ainsi, la qualité d'alignement est améliorée lorsque les mots arabes sont désagglutinés.

	Alignement1	Alignement2	Alignement3
Moyenne des probabilités de correspondance	75,98%	79,41%	82,45%
Nombre de mots anglais	3917	3917	3917
Nombre de mots arabes	3342	3071	2848

Tab. 5.2 – Résultats des trois types d'alignements après sélection des meilleures propositions

Une analyse détaillée des alignements produits conforte ces premières observations. Ainsi, au niveau des couples de mots ayant une correspondance correcte, la désagglutination des enclitiques et des proclitiques mais aussi des articles conduit à une augmentation du taux de correspondance (cf. tableau 5.3). Lorsque les correspondances ne sont pas initialement correctes (Alignement 1), la désagglutination des mots arabes permet d'obtenir des alignements de

	Alignement1	Alignement2	Alignement3
breathe	**التنفس** - 0,54 (*la respiration*)	**التنفس** - 0,65 (*la respiration*)	**تنفس** - 0,75 (*respiration*)
attack	**نوبة** - 0,33 (*crise*)	**نوبة** - 0,71 (*crise*)	**نوبة** - 0,71 (*crise*)
patients	**المرضى** - 0,4291922 (*les malades*)	**المرضى** - 0,67 (*les malades*)	**مرضى** - 1 (*malades*)

Tab. 5.3 – Amélioration des taux de correspondance correcte

meilleure qualité (cf. tableau 5.4). Grâce à la suppression des clitiques mais surtout des articles, GIZA++ propose de nouvelles correspondances proposant des mots arabes dont la sémantique est plus proche voir identique au mot anglais associé. Enfin, nous avons également observé que la désagglutination permet d'obtenir des mots arabes mieux formés notamment en supprimant les articles (cf. tableau 5.5).

	Alignement1	Alignement2	Alignement3
attacks	**والوقاية** - 0,50 (*et la prévention*)	**نوبات** - 1 (*crises*)	**نوبات** - 1 (*crises*)
shots	**بالأنسولين** - 0,33 (*avec de l'insuline*)	**الحقن** - 0,48 (*l'injection*)	**حقن** - 0,59 (*injection*)
sense	**لديهم** - 1 (*ils ont*)	**شم** - 1 (*odorat*)	**حاسة** - 1 (*sens*)
hopeless	**بالاكتئاب** - 1 (*par la dépression*)	**اليأس** - 1 (*le désespoir*)	**يأس** - 1 (*désespoir*)

Tab. 5.4 – Amélioration de la sémantique des correspondances

	Alignement1	Alignement2	Alignement3
spinach	**كالسبانخ** - 1 (*comme les épinards*)	**السبانخ** - 1 (*les épinards*)	**سبانخ** - 1 (*épinards*)
cover	**بتغطية** - 0,37 (*par couverture*)	**تغطية** - 0,67 (*couverture*)	**تغطية** - 0,63 (*couverture*)
nerves	**والأعصاب** - 1 (*et les nerfs*)	**الاعصاب** - 1 (*les nerfs*)	**اعصاب** - 1 (*nerfs*)

Tab. 5.5 – Amélioration des mots arabes alignés

Dans une deuxième étape, nous avons observé de près les différents alignements obtenus. Cette deuxième évaluation détaillée est aussi basée sur les taux de correspondance les plus élevés pour un ensemble de mots en anglais. Nous avons extrait un échantillon de 1000 mots en anglais ainsi que leurs correspondants en arabe dans chaque alignement. Cette sélection a

été effectuée au hasard. Le tableau 5.6 montre l'amélioration de la qualité d'alignement de notre échantillon en prenant en compte le phénomène d'agglutination dans les textes arabes.

	Alignement1	Alignement2	Alignement3
Nombre d'alignements corrects	660	684	692
Nombre d'alignements partiels	69	77	80
Nombre d'alignements erronés	271	239	228

Tab. 5.6 – Résultats de l'évaluation des trois types d'alignements

L'évaluation manuelle des trois alignements obtenus pour 1000 mots anglais montre que le nombre l'alignement corrects augmente par chaque étape de désagglutination passant de 660 à 692, soir une précision augmentant de 66 à 69,2%. De même, le nombre d'alignements erronés diminue de 271 à 228. En tenant compte des alignements partiellement corrects, la précision augmente de 72,9 à 76,1%.

5.3 Protocoles d'évaluation

Cette section décrit le protocole d'évaluation mis en place pour procéder à la sélection des termes candidats valides dans chaque langue ainsi que des couples de termes anglais-arabe.

5.3.1 Principes généraux

L'état de l'art montre qu'il n'existe pas de consensus bien défini pour l'évaluation des systèmes terminologiques Vivaldi and Rodríguez [2007]. Cette tâche est considérée complexe du point de vue du traitement automatique des langues. Les méthodes d'extraction terminologique sont aussi diverses que les moyens de les évaluer. Ainsi, dans le travail de Hanoka [2015], en se basant sur les travaux de Vivaldi and Rodríguez [2007], les méthodes d'évaluation des systèmes d'extraction terminologique sont classées selon les trois stratégies suivantes :

— Globale ou transparente

  — Évaluation globale : dite aussi « boite noire ». Cette méthode d'évaluation est effectuée sans détailler la démarche suivie pour avoir les résultats finaux.

  — Évaluation transparente : Cette méthode permet d'évaluer la totalité ou une partie des différents composants du processus de travail. Elle donne une explication aux erreurs obtenues et révèle les problèmes existants.

— Directe ou indirecte

  - Évaluation directe : Les méthodes d'évaluation directe consistent à déterminer la qualité des composants d'une chaîne de traitement indépendamment des différents usages qui leur sont attribués.
  - Évaluation indirecte : dite aussi « dirigée par la tâche » (*task-based*). Cette méthode permet d'évaluer les propriétés intrinsèques d'un composant d'une chaîne de traitement en déterminant l'impact de sa sortie sur la performance des autres parties de la chaîne.

- Humaine ou automatique
  - Évaluation humaine : Cette méthode consiste à évaluer manuellement la qualité des résultats obtenus. Il s'agit d'une méthode coûteuse en terme de temps. Cette évaluation peut être subjective et varie d'un évaluateur humain à un autre. Pour éviter cet inconvénient, il est nécessaire de réaliser l'évaluation deux fois, par deux évaluateurs humains différents, puis d'effectuer un consensus.
  - Évaluation automatique : Cette méthode d'évaluation repose sur une référence construite préalablement afin de la comparer aux résultats obtenus.

Pour l'évaluation des résultats obtenus lors des différentes expériences que nous avons réalisées, nous évaluons les listes de termes produites, monolingues et bilingues, suivant une méthode transparente, directe et semi-automatique : les termes candidats extraits sont confrontés à une terminologie de référence avant d'être vérifiés manuellement. Ainsi, comme ressource terminologique multilingue de référence, nous utilisons le dictionnaire multilingue en ligne ***Almaany***[1]. Ce dictionnaire comprend des mots de spécialité issus de différents domaines comme le domaine médical. La vérification humaine est effectuée par un locuteur arabe natif, non expert en médecine, mais avec des connaissances dans le domaine médical. Nous proposons un protocole d'évaluation assez détaillé en affectant pour chaque couple de termes candidats anglais-arabe cinq scores.

La présence d'un terme candidat dans le dictionnaire multilingue en ligne *Almaany* est un indice de son appartenance au domaine médical. Cependant, l'évaluation des termes complexes candidats présente des difficultés qui ont nécessité la définition de règles de validation. Ainsi, leur absence du dictionnaire n'implique pas automatiquement qu'ils ne font pas partie de la terminologie médicale pour les raisons suivantes. D'abord, comme nous l'avons mentionné à la section 3.2.1 du chapitre 3, nous traitons des textes qui proviennent des brochures à destination

1. `https://www.almaany.com/en/dict/ar-en/?c=Medical`

des patients. Etant donné que la spécialisation des termes utilisés dans le corpus peut dépendre des personnes auxquelles une terminologie est destinée (voir section 1.2.1 du chapitre 1), nous considérons ici que notre liste de termes extraits est destinée au grand public et non aux experts en médecine. De plus, chaque terme peut avoir différentes reformulations. Cependant, nos références ne couvrent pas toutes les variantes terminologiques du domaine médical. Par exemple, nous considérons que le terme candidat ***lining of the mouth*** (*muqueuses de la bouche*) correspond bien à un terme médical même s'il ne figure pas dans notre terminologie de référence. Aussi, nous avons vérifié que ses différents composants, ***lining*** (*muqueuses*) et ***mouth*** (*bouche*) font partie de cette terminologie. Ceci est le cas d'autres termes comme ***ob doctor*** (*médecin obstétricien*), ***bones of the hip joint*** (*os de l'articulation de la hanche*), ***peripheral vascular disease*** (*maladie vasculaire périphérique*), etc.

Les différents termes candidats faisant partie de notre terminologie médicale anglaise seront également évalués en se référant à la ressource en ligne ***MediLexicon***[2]. Ce portail permet aussi aux utilisateurs de rechercher les significations des acronymes et des abréviations des domaines de la médecine, de la pharmacie, de la biotechnologie, des soins de santé, etc.

5.3.2 Évaluation de termes candidats

Nous proposons d'abord une méthode d'évaluation des terminologies monolingues, c'est-à-dire des termes candidats extraits, indépendamment dans chaque langue. Nous vérifions, dans un premier temps, la structure morpho-syntaxique de chaque terme candidat pour s'assurer qu'il est bien formé. Puis, nous évaluons son appartenance au domaine médical. Pour cela, deux scores seront attribués à chaque terme candidat :

— *Forme* : nous attribuons la valeur 1 à ce score associé au terme candidat s'il constitue un terme bien formé morpho-syntaxiquement et 0 sinon.

— *Domaine* : nous attribuons la valeur 1 à ce score associé au terme candidat s'il fait partie du domaine médical et 0 sinon.

Un terme candidat est considéré donc comme un terme valide si et seulement si le produit de ses deux scores est égal à 1. Autrement dit, il doit d'abord être bien formé morpho-syntaxiquement et, évidemment, il doit faire partie du domaine médical.

2. `https://www.medilexicon.com/`

5.3.3 Évaluation de la correspondance dans un couple de termes

Le protocole appliqué aux terminologies bilingues suit une graduation à trois niveaux et se focalise sur la correspondance entre le terme de la langue source et le terme de la langue cible. Les scores présentés dans la section précédente donnent une information complémentaire sur la qualité des termes dans chaque langue. Ainsi, pour chaque couple de termes candidats, nous attribuons une lettre (*C*, *P* ou *N*) afin d'estimer la qualité de traduction entre eux :

— C : le couple de termes candidats représente une *correspondance **C**omplète*. Les deux termes candidats extraits sont correctement alignés.

— P : le couple de termes candidats représente une *correspondance **P**artielle*. Le terme candidat arabe extrait représente une partie du terme complexe arabe correspondant au terme anglais.

— N : le couple de termes candidats représente une *correspondance **N**ulle*. Il n'existe aucune correspondance entre les deux termes candidats du couple en question. Les deux termes candidats extraits ne sont pas parallèles.

5.4 Extraction monolingue : évaluation de YaTeA pour l'anglais

Dans cette section, nous évaluons la qualité de la liste des termes candidats médicaux extraits par YaTeA de nos textes en anglais. Cette liste sera utilisée par la suite lors des différents expériences visant l'apport des stratégies proposées pour l'extraction de termes arabes. Nous rappelons d'abord, dans la section 5.4.1, notre protocole d'évaluation destiné aux termes candidats monolingues anglais. Nous présentons les résultats obtenus à la section 5.4.2. Dans la section 5.4.3, nous analysons les résultats obtenus sans nécessairement chercher à relier les erreurs trouvées à des causes fines dans le fonctionnement de YaTeA puisque nous n'avons pas contribué à sa mise au point.

5.4.1 Rappel du protocole d'évaluation

Avant de présenter les résultats obtenus par YaTeA, nous rappelons brièvement du protocole suivi pour l'évaluation des termes candidats anglais (cf. section 5.3.2). Deux scores sont attribués à chaque terme candidat extrait. Le premier score reflète sa structure morpho-syntaxique : la valeur de 1 est attribuée s'il s'agit d'un terme bien formé morpho-syntaxiquement et 0 sinon. Le

deuxième score reflète l'adéquation du terme candidat extrait au domaine médical : la valeur de 1 ou 0 lui est attribuée s'il s'agit ou non d'un terme médical. Pour cela, nous nous appuyons sur les terminologies médicales anglaises issues de *Almaany* et *MediLexicon*.

5.4.2 Expérience et résultats

Les méthodes d'extraction de termes arabes par translittération (section 4.3) et par transfert (section 4.4) nécessitent de disposer d'une liste de termes issus de la langue source, dans notre cas, l'anglais. La qualité des termes anglais pouvant avoir un impact sur les résultats de l'extraction de termes arabes, nous avons analysé cette liste de termes candidats extraits automatiquement par YaTeA.

Pour cela, nous avons donc utilisé YaTeA sur le corpus anglais décrit à la section 3.2. Nous avons obtenu 4634 termes candidats anglais dont la majorité, soit 59,6%, représente des termes candidats complexes (cf. tableau 5.7).

Nombre total des termes candidats	4634
Nombre de termes candidats simples	1872 (40,4%)
Nombre de termes candidats complexes	2762 (59,6%)

Tab. 5.7 – Résultats de l'extraction terminologique pour l'anglais

Nous avons analysé les 500 premiers termes candidat extraits par YaTeA afin d'effectuer une évaluation semi-automatique en s'appuyant sur le protocole défini à la section 5.4.1. A partir de notre échantillon de 500 termes candidats anglais extraits, 385 (soit 77%) sont considérés comme termes médicaux corrects, bien formés morpho-syntaxiquement et faisant partie au domaine médical. Cependant, 115 (soit 23%) sont rejetés. Le produit de leurs scores est égal à zéro (voir 5.3.2). Le tableau 5.8 présente les différents résultats obtenus.

Nombre de termes candidats évalués	500
Nombre de termes candidats validés	385 (soit 77%)
Nombre de termes candidats rejetés	115 (soit 23%)

Tab. 5.8 – Résultats de l'extraction terminologique pour l'anglais sur les 500 premiers termes candidats extraits

5.4.3 Analyse des erreurs

Nous avons observé que la plupart des termes candidats rejeté pendant l'étape d'évaluation ne font pas partie du domaine médical bien qu'il sont bien formés morpho-syntaxiquement. Par exemple, les termes candidats ***air bubbles***, ***minutes*** et ***few seconds*** figurent dans la liste extraite par YaTeA mais ils n'appartiennent pas au domaine médical.

D'autres termes sont rejetés bien qu'ils se rapportent au domaine médical car ils ne représentent pas des termes candidats bien formés morpho-syntaxiquement. Ceci est dû aux erreurs d'étiquetage morpho-syntaxique réalisé par l'étiqueteur TreeTagger. Par exemple, le mot ***talk*** dans le terme candidat ***talk to the staff*** est considéré comme un *nom* et non comme *verbe*.

5.5 Extraction monolingue : évaluation de l'adaptation de YaTeA à l'arabe

Dans cette section, nous évaluons les termes candidats extraits par l'outil YaTeA adapté pour la langue arabe. Nous rappelons d'abord le protocole d'évaluation suivi spécifiquement pour les termes monolingues (section 5.5.1), puis nous décrivons les deux expériences réalisées, qui prennent en compte ou non les phénomènes d'agglutination dans le processus d'extraction de termes (section 5.5.2). Les résultats des différentes expériences sont présentés dans la section 5.5.2. puis discutés dans la section 5.5.3.

5.5.1 Rappel du protocole d'évaluation

Pour estimer la qualité des termes candidats proposés par cette première stratégie d'acquisition terminologique pour l'arabe, nous avons suivi le protocole spécifique à l'évaluation des termes candidats monolingues, décrit dans la section 5.3.2. Ainsi, nous avons procédé à une évaluation manuelle en se référant à la terminologie médicale arabe proposée par *Almaany*, et nous considérons qu'un terme candidat est correct s'il est bien formé morpho-syntaxiquement et s'il appartient au domaine médical.

5.5.2 Expériences et résultats

Nous avons évalué l'adaptation de l'extracteur de termes YaTeA sur les 30 textes médicaux arabes que nous avions préparés et nettoyés au début de la thèse. Cela représente un corpus de 15 532 mots non désagglutinés.

A travers deux expériences, nous avons voulu identifier l'impact de la prise en compte de l'agglutination dans le processus d'extraction de termes. Les résultats de l'extraction de termes obtenus lors de ces deux expériences sont présentés dans le tableau 5.9.

Dans chaque expérience, les deux étapes du processus d'extraction de termes seront évaluées. La première étape correspond au découpage du corpus à l'aide de frontières syntaxiques. Elle permet d'obtenir des syntagmes nominaux maximaux pouvant constituer ou contenir des termes candidats. La deuxième étape permet de produire des termes candidats complexes, mais aussi des termes candidats simples. Ces termes candidats sont extraits suite à l'application des patrons syntaxiques.

	Étape 1		Étape 2			
	SNM	TS	TS	TCmax	TC	Total
Pas de prise en compte de l'agglutination	1972	262	262	590	1133	1395
Prise en compte de l'agglutination	1916	298	298	400	824	1122

Tab. 5.9 – Résultats de l'extraction de termes sur les textes médicaux en arabe (SNM : syntagmes nominaux maximaux, TS : termes simples candidats, TCmax : termes complexes candidats correspondant aux syntagmes nominaux maximaux, TC : termes complexes candidats).

Expérience 1 – non prise en compte de l'agglutination Dans un premier temps, nous avons utilisé l'extracteur de termes adapté pour l'arabe sans prendre en compte le phénomène d'agglutination. Les enclitiques et les proclitiques sont alors considérés comme faisant partie des mots.

Comme présenté dans le tableau 5.9, 1972 syntagmes nominaux maximaux (SNM) ainsi que 262 noms (TS) qui sont considérés comme des termes simples candidats, sont extraits grâce aux frontières syntaxiques définies pour l'étape 1. L'analyse syntaxique des syntagmes nominaux maximaux (étape 2) permet de retenir 590 syntagmes nominaux maximaux (TCmax). Un terme complexe pouvant contenir des termes simples ou complexes, les constituants des termes complexes sont également considérés comme des termes candidats. Nous disposons donc d'un ensemble de 1395 termes candidats dont 1133 termes candidats complexes (TC). Par exemple, le terme candidat médical arabe **ارتفاع معدلات الكولسترول** (*élévation du taux de cholestérol*) inclut à la fois le terme candidat complexe **معدلات الكولسترول** (*taux de cholestérol*) et les termes simples **ارتفاع** (*élévation*), **معدلات** (*taux*) et **الكولسترول** (*le cholestérol*).

Nous avons analysé manuellement les 262 termes simples (TS) et les 590 termes candidats correspondant aux syntagmes nominaux maximaux (TCmax). Grâce à cette validation, nous avons pu évaluer la qualité de l'analyse syntaxique et la pertinence des termes candidats extraits. Ainsi, parmi les 590 syntagmes nominaux maximaux, 388 termes candidats (65,7%) sont jugés correctement analysés et pertinent pour le domaine médical. Il ressort de cette analyse que l'agglutination, et en particulier les proclitiques, est la principale source d'erreurs aussi bien lors de l'utilisation des frontières syntaxiques que de l'analyse syntaxique des syntagmes.

Expérience 2 – prise en compte de l'agglutination La deuxième expérience a pour objectif d'évaluer la contribution des traitements spécifiques visant à séparer une classe des proclitiques, les prépositions, des mots auxquels ils sont associés. La première étape permet d'extraire 298 termes candidats simples (TS) et 1916 syntagmes nominaux maximaux (SNM). Parmi ces derniers, 400 sont conservés à la fin de l'étape 2 et permettent d'obtenir 824 termes complexes candidats. Nous avons également analysé les 400 syntagmes nominaux maximaux retenus : 288 (72,1%) sont jugés corrects. La figure 5.1 présente quelques termes extraits des textes médicaux arabes.

La prise en compte de l'agglutination se caractérise par un nombre moins élevé de termes complexes candidats extraits et une augmentation du nombre de termes simples. Cette augmentation peut s'expliquer par le fait que des éléments initialement agglutinés (articles ou prépositions) sont alors considérés comme des frontières syntaxiques. Nous observons également que le nombre de syntagmes nominaux maximaux non analysés syntaxiquement augmente fortement. Nous expliquons cela par le regroupement, plus ou moins fortuit, des mots privés des proclitiques associés dans des syntagmes maximaux plus grands et, par conséquent, plus difficilement analysables ou ne correspondant pas des termes candidats syntaxiquement bien formés. Une analyse approfondie des syntagmes maximaux non analysés permet de confirmer cette hypothèse.

سرطان الثدي	(cancer du sein)	سرعة ضربات القلب	(rythme cardiaque rapide)
الاوعية الدموية	(les vaisseaux sanguins)	الذراع الايمن	(le bras droit)
تمارين الكاحلين	(exercices des chevilles)	درجة الالم	(degré de la douleur)
الزائدة الدودية	(l'appendice)	العلاج الكيماوي	(chimiothérapie)
العلاج الاشعاعي	(la radiothérapie)	اخذ عينات انسجة الثدي	(biopsies mammaires)
مرض السكر	(le diabète)	الرئتين	(les deux poumons)
التهاب شعب هوائية	(bronchite)	آلام	(douleurs)
العلاج	(le traitement)	الرأس	(la tête)

Fig. 5.1 – Exemple de termes extraits en MSA

5.5.3 Analyse des erreurs

Dans cette section, nous exposons les différentes raisons qui ont empêché l'extraction des termes aussi bien lors de l'étape 1 que de l'étape 2. Un premier type d'erreurs est lié à la qualité de l'étiquetage morpho-syntaxique de MADA+TOKAN. Certains termes n'ont pas pu être analysés et identifiés comme des termes car les étiquettes morpho-syntaxiques associées aux mots sont erronées. Par exemple, pour le terme **العضلة رباعية الرؤوس** (*litt. le muscle à quatre têtes*) (*quadriceps*), le mot **العضلة** (*le muscle*) est considéré comme un adjectif alors qu'il s'agit bien d'un nom et le mot **الرؤوس** (*les têtes*) est considéré comme un nom masculin singulier alors qu'il est au pluriel.

Par ailleurs, comme dans toutes les langues, des erreurs d'étiquetage sont dues aux mots inconnus ou aux termes étrangers empruntés pour le MSA et translittérés en caractères arabes. Le terme **الفولات** (*acide folique*) illustre ce cas de figure. Celui-ci est considéré comme un nom féminin pluriel pouvant être traduit par le mot *les fèves*.

La non-voyellation des textes arabes introduit des ambiguïtés dans les formes fléchies, ce qui peut conduire à un mauvais étiquetage morpho-syntaxique. Ainsi, le mot non-voyellé **تناول** peut correspondre à la forme verbale (*a pris*) ou au nom (*la prise*). Sa catégorisation comme verbe conduit à rejeter le terme **تناول الفيتامينات** (*prise des vitamines*).

Enfin, le phénomène d'agglutination a également un impact sur la qualité de l'étiquetage morpho-syntaxique. Ces erreurs empêchent l'extraction de termes médicaux arabes. Par exemple, le mot **بثني** (*par flexion*) est considéré comme une expression verbale (*il m'a diffusé*), alors qu'il devrait être décomposé en un nom **ثني** (*flexion*) précédé par le proclitique **ب** (*par*). Il n'est alors pas possible d'extraire les termes complexes incluant ce mot, comme par exemple **ثني الركبة** (*flexion du genou*).

Le deuxième type d'erreur est dû au manque des patrons syntaxiques adaptés aux différents structures des termes.

5.6 Évaluation des termes arabes extraits par translittération

Dans cette section, nous évaluons et discutons notre méthode d'extraction de termes arabes basée sur la translittération des termes anglais en caractères arabes (voir section 4.3). Nous rappelons le protocole d'évaluation et nous détaillons ses spécificités pour cette méthode dans la section 5.6.1. La section 5.6.2 expose les expériences et résultats obtenus. Finalement, nous

présentons une analyse des erreurs à la section 5.6.3.

5.6.1 Rappel du protocole d'évaluation

Notre objectif est d'évaluer précisément notre méthode d'extraction des termes arabes translittérés à partir de termes anglais, et par effet de bord, les couples de termes anglais-arabe. Il s'agit aussi de connaître l'origine des erreurs produites par la méthode proposée et de mieux identifier ses limites. Aussi, nous vérifions, pour chaque couple de termes candidats, les aspects suivants :

— la qualité des termes candidats anglaise et arabe extraits ;

— la translittération du terme candidat anglais en caractères arabes ;

— la correspondance des deux termes candidats anglais et arabe.

Pour vérifier s'il s'agit bien de termes médicaux, nous avons utilisé, d'une part, le dictionnaire *Almaany* et en particulier, sa terminologie médicale multilingue en ligne, et d'autre part, la base de données médicale anglaise en ligne *MediLexicon*.

Nous rappelons que l'évaluation de la qualité de translittération des termes candidats extraits a été effectuée manuellement car un terme anglais translittéré en caractères arabes peut avoir différentes formes (voir section 4.3.3 du chapitre 4). Lors de cette étape d'évaluation, nous admettons les règles suivants :

— Un terme candidat anglais sera considéré comme terme médical s'il figure dans l'une de nos références.

— Si le terme candidat anglais est bien un terme médical et qu'il est correctement translittéré, le terme candidat arabe sera considéré comme son parallèle en MSA.

— Les couples de termes candidats ayant une translittération correcte forment un bon alignement.

5.6.2 Expériences et résultats

Nous avons utilisé 12 textes parallèles anglais-arabe pour l'étude du corpus et pour la définition de notre méthode d'extraction de termes arabes par translittération, c'est-à-dire de termes anglais translittérés en caractères arabes. Pour évaluer notre méthode, nous avons utilisé 90 textes parallèles anglais-arabe supplémentaires. L'alignement étant réalisé de manière non-supervisée, nous avons utilisé l'ensemble des textes du corpus pour aligner les textes au niveau des mots. Les

caractéristiques de notre corpus constitué de 102 textes parallèles anglais-arabe sont présentées dans le tableau 5.10.

	Corpus d'étude		Corpus de test	
	Corpus anglais	Corpus arabe	Corpus anglais	Corpus arabe
Nombre de textes	12	12	90	90
Nombre de mots	5692	7699	55208	72204

Tab. 5.10 – Caractéristiques du corpus

L'application de la méthode d'extraction de termes par translittération permet d'obtenir 137 termes arabes, chaque terme étant issu d'un seul couple de termes anglais-arabe. L'évaluation de ces 137 termes candidats arabes est réalisée semi-automatiquement suivant le protocole décrit ci-dessus. Le tableau 5.11 récapitule les caractéristiques des résultats obtenus par notre méthode d'extraction des termes arabe. La méthode permet d'obtenir 116 termes candidats ayant une translittération correcte (soit 84,67%) dont 92 sont des termes médicaux (soit 79,31% des translittérations correctes et 67,15% des résultats).

En outre, parmi les translittérations erronées nous constatons que dans 8,03% des termes candidats anglais sont bien alignés avec des termes candidats arabes mais ne permettent pas d'identifier une relation de translittération entre les termes candidats anglais et arabe. Par exemple, le terme candidat anglais ***calcium-fortified*** (*enrichi en calcium*) est associé au terme candidat arabe **كالسيوم** (*calcium*) alors que ce dernier représente une partie du terme candidat anglais.

Même si la translittération d'un terme candidat simple anglais est correcte, ceci n'implique pas forcément qu'il s'agit d'un terme médical. Parmi la liste extraite, il existe des mots étrangers empruntés pour l'arabe. C'est le cas des noms correspondant à des cognats comme par exemple le couple ***Ohio*** - **اوهايو** qui désigne le nom d'un État. De même pour le couple ***Columbus*** - **كولومبوس** qui désigne une ville.

Par ailleurs, une minorité des termes candidats extraits (7,3%) n'est ni alignée ni translittérée. Ils sont dus principalement aux erreurs produites à partir de l'alignement fourni par GIZA++. Par exemple, le terme candidat anglais ***grounds*** est aligné avec un point *".*", signe de ponctuation. La figure 5.2 présente des exemples de couples de termes anglais-arabe extraits par translittération.

Les noms des médicaments font aussi partie de la terminologie médicale et ils figurent dans les terminologies de référence issues de *Almaany* et *MediLexicon*. Nous avons ainsi observé que 9 noms de médicaments figuraient parmi les 101 termes translittérés.

بنكرياس	*pancreas*	بكتريا	*bacteria*
كريم	*cream*	بروتين	*protein*
تيتانوس	*tetanus*	باريوم	*barium*
باكسيل	*paxil*	نويرونتين	*neurontin*
فولات	*folate*	ترمومتر	*thermometer*

Fig. 5.2 – Exemple de termes translittérés extraits

Nombre de termes/couples extraits			
137			
Translittérations correctes		Translittérations erronées	
116 (84,67%)		21 (15,33%)	
Termes	Non Termes	Alignement correct	Alignement non correct
101 (73,72%)	15 (10,95%)	11 (8,03%)	10 (7,3%)

Tab. 5.11 – Résultats de l'extraction des termes arabes par translittération (le nombre de couples est identique au nombre de termes car chaque terme candidat arabe est issu d'un seul couple de termes)

5.6.3 Analyse des erreurs

L'analyse des résultats conduit à plusieurs observations et permet d'identifier des limites à la méthode proposée. Nous les détaillons dans cette section.

Tout d'abord, comme il s'agit d'un alignement au niveau des mots, certains termes complexes sont alignés avec une partie du terme arabe qui devrait correspondre au terme anglais en question. Ainsi, par exemple, nous constatons que le terme ***x-ray*** (*rayon-X*) est aligné avec le terme simple arabe **اشعة** (*rayons*) qui représente une partie du terme complexe arabe **اشعة-اكس** (*rayon-X*). Dans le corpus, le trait d'union n'est pas utilisé pour lier les deux composants du terme arabe (**اكس** et **اشعة**). L'alignement n'est alors réalisé qu'avec un seul des deux composants et empêche l'extraction par translittération du terme arabe correspondant au terme anglais ***x-ray***.

La qualité de l'alignement joue un rôle important dans les résultats obtenus. Ainsi certains termes arabes translittérés ne figurent pas dans la liste produite par notre méthode car les mots les composant sont mal alignés. Par exemple, le terme arabe **زنك** (*zinc*) est aligné avec le terme ***oxide*** alors que **اكسيد** (*oxide*) est aligné avec le terme ***zinc***.

Nous avons également observé que les textes spécialisés anglais, mais aussi les ressources terminologiques, emploient des acronymes pour désigner des concepts, alors que les textes arabes utilisent les termes correspondant. Cet usage des acronymes influe sur les résultats de la méthode proposée. Par exemple, l'acronyme ***MRSA*** (*SARM*), correspondant au terme ***Methicillin-Resistant Staphylococcus Aureus*** (*Staphylococcus aureus résistant à la méticil-*

line), est aligné avec le mot arabe **للمثيسيلين**[3] (*à la méticilline*). Le terme candidat arabe extrait (**للمثيسيلين**) représente alors une partie du terme arabe complexe **العنقوديات الذهبية المقاومة للمثيسيلين** correspondant au terme anglais ***MRSA***.

De manière similaire, les terminologies et textes spécialisés anglais se caractérisent par l'utilisation d'abréviations. Ainsi, comme pour les acronymes, certains couples de termes dont la partie arabe représente une translittération ne peuvent être extraits par notre méthode. Contrairement aux acronymes où l'alignement posait problème, ici, la difficulté était lié à la translittération. Ainsi, notre méthode conduit à translittérer une abréviation anglaise en arabe alors que c'est la forme non-abrégée correspondante qui est translitérée en arabe. Par exemple, le couple de termes ***flu*** et **انفلونزا** (*grippe*), c'est-à-dire la translittération de ***influenza***, ne peut être proposé par une méthode basée sur la translittération car le terme anglais ***influenza*** est représenté par son abréviation *flu*.

Comme mentionné précédemment, la terminologie médicale arabe repose sur l'assimilation des termes d'origine étrangère. L'origine des termes translittérés a donc une influence sur les termes extraits. Les résultats obtenus avec la méthode que nous avons proposée, dépendent des langues des corpus utilisés pour nos expériences. Ainsi, si nous avions utilisé un corpus parallèle français-arabe, d'autres termes auraient été extraits par translittération. Par exemple, le terme arabe **بوصة** est une translittération du terme français *pouce*. Cette observation explique l'absence du terme arabe *inch* - **بوصة** (*pouce*) dans les résultats que nous obtenons sur le corpus parallèle anglais-arabe. De même, les termes ***iodine*** et **يود** (*iode*) sont considérés comme correctement alignés mais le terme **يود** n'est pas la translittération du terme ***iodine***, car, bien que ce terme arabe soit emprunté à une langue étrangère, il s'agit de la translittération du terme français ***iode*** en caractères arabes. L'application de la méthode d'extraction de termes arabes par translittération sur des corpus parallèle spécialisés dans différentes langues devrait ainsi permettre d'améliorer la couverture de la terminologie construite automatiquement.

L'analyse des résultats obtenus montre également qu'une terminologie arabe peut également contenir des termes anglais translittérés bien qu'il existe déjà un terme arabe correspondant. Autrement dit, un terme anglais peut à la fois avoir comme correspondant sa translittération en caractères arabes ainsi qu'un autre terme arabe ayant le même sens, les deux termes arabes pouvant être considérés comme des synonymes. Par exemple, le terme *ounces* - (onces) peut correspondre au terme arabe **اوقيات** ou à sa translittération, c'est-à-dire le terme **اونصة**.

3. ce terme arabe n'est pas désagglutiné car il s'agit d'un mot étranger que MADAMIRA n'arrive pas à identifier.

5.7 Évaluation de l'extraction terminologique par transfert anglais-arabe

Dans cette section, nous évaluons la méthode d'acquisition terminologique basée sur la notion de transfert translingue proposée à la section 4.4. Les résultats sont obtenus par l'application de cette méthode sur le corpus parallèle anglais-arabe décrit au chapitre 3. D'abord, nous rappelons le protocole d'évaluation suivi et ses spécificités pour les données extraites dans la section 5.7.1. Par la suite, nous présentons les expériences et les résultats obtenus dans la section 5.7.2. Finalement, nous discutons ces résultats à la section 5.7.3.

5.7.1 Rappel du protocole d'évaluation

Pour l'évaluation de la méthode d'extraction des termes médicaux arabes par transfert, nous avons suivi le protocole complet présenté à la section 5.3. Dans un premier temps, nous avons évalué les termes candidats anglais et arabes. Pour cela, nous avons vérifié la forme morpho-syntaxique du terme candidat ainsi que son appartenance au domaine médical en nous référant aux dictionnaires multilingues en ligne *Almaany* et *MediLexicon*.

Comme dans toutes les langues, les termes figurent généralement dans les terminologies arabes au nominatif. Cependant, comme il s'agit de l'extraction des termes candidats arabes parallèles, nous avons considéré les termes indépendamment des marques morphologiques de cas.

De même, la traduction des textes de l'anglais en arabe peut être inexacte ou refléter une notion plus générale. Nous avons choisi de considérer que le terme arabe extrait par transfert est pertinent. Par exemple, le syntagme nominal ***ankle pumps*** (*pompes à la cheville*) présente un terme candidat anglais extrait par YATEA, désignant un exercice pour réduire les maux des chevilles. Comme notre corpus est dédiées aux patients, ce terme candidat est considéré comme étant un terme médical anglais. De plus, le mot ***ankle*** (*cheville*) figure dans *Almaany* comme terme médical. Par contre, ***pumps*** (*pompes*) y apparaît comme composant pour d'autres termes médicaux complexes. Cependant, la phrase en arabe contient le terme candidat **تمارين الكاحلين** (*exercices des chevilles*). Bien qu'il ne s'agisse pas de la traduction exacte du terme anglais mais plutôt d'un concept plus général du terme anglais en question, nous l'avons retenu comme un terme arabe pertinent.

Dans un deuxième temps, nous avons évalué le degré de correspondance entre les termes candidats anglais et arabe comme décrit à la section 5.3.3 : correspondance complète, partielle

ou pas de correspondance.

5.7.2 Expériences et résultats

Nous avons évalué notre méthode d'acquisition des termes médicaux arabes par transfert sur un corpus de 102 textes parallèles anglais-arabe composé de 60900 mots anglais et 79903 mots arabes. Le tableau 5.12 récapitule les caractéristiques de notre corpus d'évaluation utilisé. Nous rappelons que notre corpus arabe est désagglutiné au niveau des enclitiques et les proclitiques, y compris les articles.

Nombre de textes	102
Nombre de mots arabes	79903
Nombre de mots anglais	60900

Tab. 5.12 – Caractéristiques du corpus d'évaluation

L'application de la méthode d'acquisition des termes arabes par transfert permet d'obtenir 4963 termes arabes. Nous considérons que les termes candidat arabes ayant la même forme lemmatisée représente un seul terme arabe candidat. Nous avons obtenu 6994 combinaisons anglais-arabe des termes arabes aux formes fléchies correspondant à 6552 couples de termes candidats anglais-arabe. A partir de ces couples de termes candidats, 3128 termes anglais candidats ont été repérés. Le tableau 5.13 présente les résultats obtenus par notre méthode d'extraction de termes arabes par transfert.

Nombre des termes candidats arabe	4963
Nombre des couples de termes candidat anglais-arabe	6994
Nombre des termes candidat anglais	3128

Tab. 5.13 – Résultats de l'acquisition des termes candidats arabes par transfert

Nous avons réalisé des évaluations sur un échantillon de 1000 couples de termes candidats tirés au hasard sur 6994 couples au total. Ces couples correspondent à 952 termes candidats arabes et 448 termes candidats anglais. Cette évaluation a été effectuée semi-automatiquement en s'appuyant sur le protocole d'évaluation présenté à la section 5.3.2. Deux scores seront attribués pour chacun des termes afin d'estimer sa forme morpho-syntaxique ainsi que son appartenance au domaine médical. Le tableau 5.14 présente les caractéristiques des termes arabes extraits. Parmi les 952 termes candidats arabes extraits, 354 (soit 37,6%) ont été validés comme termes médicaux arabes bien formés. Ceux-ci représentent les termes candidats ayant le produit des deux score égal à 1. Cependant, 598 termes candidats, (soit 62,4%) ont été rejetés. Une analyse

plus détaillée montre que 184 d'entre eux (soit 30,77% des termes candidats rejetés et 19,33% de la totalité) représentent des termes candidats qui sont bien formés morpho-syntaxiquement mais qui ne font pas partie du domaine médical. Aussi, 95 termes candidats arabes extraits (soit 15,87% des termes candidats rejetés et 9,98% de la totalité des termes candidats extraits) font partie du domaine médical mais ne sont pas bien formés morpho-syntaxiquement.

	Médical	Non médical	Total
Bien formé	358 (37,6%)	184 (19,33%)	542 (56,93%)
Mal formé	95 (9,98%)	315 (33,09%)	410 (43,07%)
Total	453 (47,58%)	499 (52,42%)	952 (100%)

Tab. 5.14 – Caractéristiques des termes candidats arabes extraits

Nous avons aussi étudié la qualité des couples de termes candidats anglais-arabe selon notre protocole d'évaluation présenté à la section 5.3.3. Pour chacun de ces couples, nous attribuons la lettre *C* (*correspondance **C**omplète*), *P* (*correspondance **P**artielle*) ou *N* (*correspondance **N**ulle*) afin d'estimer le lien entre le terme candidat arabe et celui en anglais. Le tableau 5.15 présente les résultats obtenus selon les différents types de correspondances obtenues. Pour chacune d'entre elles, nous identifions le nombre des couples de termes candidats anglais-arabe qu'elle comporte, ainsi que le nombre des termes candidats anglais sans doublon.

Type de correspondance	Nb. couples de termes candidats	Nb. termes candidats anglais
Correspondance complète	552	309
Correspondance partielle	129	112
Pas de correspondance	319	169
Total	1000	590

Tab. 5.15 – Résultats de l'acquisition par transfert selon le type de correspondance obtenue, sur l'échantillon de 1000 couples de termes

Parmi les 552 couples de termes candidats anglais-arabe ayant une correspondance complète, c'est-à-dire le terme arabe est la traduction du terme anglais, 296 couples (soit 53,62%) comportent des termes anglais du domaine médical. Parmi ceux-ci, 284 couples (soit 51,45%) sont considérés comme couples de termes médicaux anglais-arabe. Le tableau 5.16 détaille ces résultats.

Sur l'ensemble de 1000 couples de l'échantillon, 129 couples de termes candidats ont une correspondance partielle, c'est-à-dire le terme arabe représente une partie de la notion véhiculée par le terme anglais. Parmi ces couples, 101 (78,29%) comportent des termes médicaux anglais dont 36 (27,91%) sont associés à des termes médicaux arabe même s'ils ne représentent pas des

	Termes arabes	Non-termes arabes	Total
Termes anglais	284 (51,45%)	12 (2,17%)	296 (53,62%)
Non-termes anglais	0 (0,0%)	256 (46,38%)	256 (46,38%)
Total	284 (51,45%)	268 (48,55%)	552 (100%)

Tab. 5.16 – Répartition des 552 couples de termes ayant une correspondance complète, issus de l'échantillon de 1000 couples de termes

correspondances partielles. Le tableau 5.17 présente les couples ayant ce type de correspondance.

	Termes arabes	Non-Termes arabes	Total
Termes anglais	36 (27.91%)	65 (50.39%)	101 (78.3%)
Non-Termes anglais	0 (0.0%)	28 (21.70%)	28 (21.70%)
Total	36 (27.91%)	93 (72.09%)	129 (100%)

Tab. 5.17 – Répartition des 129 couples des termes ayant une correspondance partielle, issus de l'échantillon de 1000 couples de termes

Parmi les couples analysés, 319 couples de termes n'impliquent pas une relation de traduction entre le terme anglais et le terme arabe. Cependant, nous avons observé que ces couples comportent 149 termes médicaux anglais (soit 46,71%) et 47 termes médicaux arabes (soit 14,73%) n'ayant pas de lien entre eux. Le tableau 5.17 présente la composition des couples de termes n'ayant pas de correspondance.

	Termes arabes	Non-Termes arabes	Total
Termes anglais	27 (8.46%)	122 (38.24%)	149 (46.71%)
Non-Termes anglais	20 (6.27%)	150 (47.02%)	170 (53.29%)
Total	47 (14.73%)	272 (85.27%)	319 (100%)

Tab. 5.18 – Répartition des 319 couples des termes n'ayant pas de correspondances, issus de l'échantillon de 1000 couples de termes

Dans une troisième étape, nous avons évalué les termes candidats anglais extraits. Nous avons validé 292 termes anglais parmi les 448 termes candidats extraits, soit 65,18%. L'extraction terminologique monolingue à partir des textes anglais est plus détaillée à la section 5.4. Ceci montre que les erreurs de l'extracteur des termes YATEA pour la langue anglaise est l'une des sources d'erreurs lors de l'extraction de termes arabes par transfert. La figure 5.3 présente des exemples des couples de termes anglais-arabe extraits. Nous remarquons qu'un terme arabe peut correspondre à un ou plusieurs termes anglais, et inversement, un terme anglais peut correspondre à un ou plusieurs termes arabes. Nous constatons aussi la présence de certains termes arabes translittérés de l'anglais comme **كولسترول** (*cholestérol*).

تمزق جلدي	*skin lesion*	عظم	*bone*
قسم الطوارئ	*emergency department*	مؤشرا علي نوبة قلبية	*sign of a heart attack*
جراحة	*surgery*	جانب الخنصر	*little finger side*
اكسيد الزنك	*zinc oxide*	نويرونتين	*neurontin*
طبية	*medical*	اوعية دموية	*blood vessels*
ارتفاع الكوليستيرول	*cholesterol*	ارتفاع معدلات الكولسترول	*high cholesterol*
كولسترول	*cholesterol*	باريوم	*barium*
مشكلات في التنفس	*trouble breathing*	مشكلات في التنفس	*problems breathing*

Fig. 5.3 – Exemple de couples de termes anglais arabes extraits par transfert

5.7.3 Analyse des erreurs

Un nombre important des termes candidats arabes extraits rejetés provient de l'absence de correspondance entre les termes candidats anglais et ceux en arabes. Plusieurs couples de termes candidats comportent un terme anglais dont le terme arabe associé représente une partie du terme arabe complexe attendu. Ceci est dû principalement aux erreurs produites par la phase d'alignement, en particulier lorsqu'il s'agit de termes anglais simples. Par exemple, le terme candidat anglais ***sunscreen*** est aligné avec **واقي** (*protecteur*) ou **شمس** (*soleil*), une partie du terme arabe **واقي شمس** (*litt. protecteur du soleil*) (*écran solaire*) ou encore, le terme ***decaf*** (*décaféiné*) est aligné avec **منزوعة** (*enlevé*) une partie du terme arabe **منزوعة الكافيين** (*litt. caféine enlevé*). De plus, certaines caractéristiques de la langue médicale anglaise comme les termes anglais simples formés grâce à la composition morphologique et l'usage des acronymes accentuent ce type de problème.

Termes anglais composés morphologiques Nous avons observé que plusieurs des termes médicaux anglais simples, essentiellement des nom des traitements et des maladies, sont associés à des termes arabes complexes. Un nombre important de ces termes simples anglais sont des composés morphologiques. Chacun des composants de ces termes doit être traduit en arabe sous forme d'un mot ou d'un terme. Les termes arabes correspondant attendus sont donc des termes complexes. Il en résulte une difficulté pour notre méthode à extraire le terme arabe : si le terme anglais est simple, l'alignement correct doit associer plusieurs mots arabes au mot anglais. Par exemple le terme ***Appendectomy*** (*appendicectomie*) est un terme anglais simple morphologiquement composé. Il peut être décomposé en deux parties ; i) le composant ***append*** renvoyant au terme ***appendix*** (*appendice*) et ii) le composant ***-ectomy*** (*-ectomie*) désignant une opération chirurgicale ou l'enlèvement d'une partie d'un organe. En arabe, ce terme correspond au terme complexe suivant **استئصال الزائدة الدودية** dans lequel **استئصال** (*ablation/enlèvement*) correspond au composant ***-ectomy*** et **الزائدة الدودية** (*appendice*) correspond au composant ***append***.

De même, le terme ***immunodeficiency*** (*immunodéficience*) est la combinaison des deux composants ***immune*** (*immunité*) et ***deficiency*** (*déficience*). Il correspond au terme arabe complexe **نقص المناعة** dont le terme **نقص**(*manque*) représente le composant ***deficiency*** (*déficience*) ou (*déficit*) et le terme **المناعة** (*l'immunité*) représente le composant ***immune***.

Acronymes Comme indiqué dans la section 5.6.3, l'usage d'acronyme est commun dans les textes de spécialités et les terminologies en anglais. Cependant, cet usage n'est pas possible en arabe. Il est alors nécessaire d'utiliser la traduction de la forme développée de l'acronyme. Ce phénomène est également une source d'erreurs dans notre approche par transfert, où comme précédemment, la qualité de l'alignement joue un rôle important. Ainsi, par exemple, le terme ***UTI*** qui représente l'acronyme du terme complexe *'**U**rinary **T**ract **I**nfection* (*infection des voies urinaires*), doit être aligné avec le terme complexe arabe : **التهاب المسالك البولية** (**التهاب** - ***inflammation*** (*inflammation*), **المسالك** - ***tract*** (*des voies*), **البولية** - ***urinary*** (*urinaires*)).

5.8 Bilan

Dans ce chapitre, nous avons évalué les différentes méthodes proposées pour l'extraction des termes. Pour cela, nous avons utilisé des textes médicaux arabes et analysé les résultats obtenus. Compte tenu des résultats obtenus, nous envisageons d'extraire les termes translittérés à partir de phrases parallèles plutôt qu'à partir des couples de termes candidats alignés anglais-arabe, pour éviter que cette méthode dépende de la qualité de l'alignement au niveau des mots. De plus, nous proposons d'intégrer des patrons syntaxiques définis lors de l'adaptation de Y_ATEA pour le MSA pour corriger les termes candidats arabes mal formés qui sont extraits par le transfert et font partie du domaine médical.

Nous revenons maintenant à la question de recherche abordée dans cette thèse : lorsqu'il n'existe pas d'outil pour réaliser une tâche donnée sur une langue, est-il préférable i) d'adapter un outil existant, ii) de mettre en oeuvre une méthode de transfert translingue, ou iii) de mettre au point une méthode spécifique monolingue.

Les résultats obtenus et leur évaluation montrent que l'adaptation d'un extracteur de termes (Y_ATEA) pour le MSA produit plus de termes bien formés morpho-syntaxiquement. Mais la qualité de l'étiquetage morpho-syntaxique et le manque de patrons adaptés aux différentes structures des termes arabes influent sur les résultats obtenus et nécessite un long travail de description du processus d'extraction terminologique.

Par contre, nous ne rencontrons pas cet inconvénient avec la méthode d'acquisition des

termes par transfert. Elle n'exige pas une étude préalable approfondie sur la composition des termes arabes. Mais ici, l'alignement de notre corpus parallèle au niveau des mots joue un rôle très important dans la qualité des termes candidats extraits. Celle-ci dépend également de l'extraction monolingue effectuée à partir des textes anglais.

Enfin, les deux contraintes évoquées ci-dessus influent aussi sur notre méthode d'extraction des termes médicaux anglais translittérés en caractères arabes, celle-ci étant limitée à l'extraction de termes simples.

Chapitre 6

Conclusion

Sommaire

6.1 Bilan

Dans les domaines de spécialité tels que le domaine médical, l'accès à l'information présente dans les textes nécessite de disposer de ressources terminologiques. Ces ressources recensent les termes, en général des groupes nominaux, représentant les notions du domaine. Dans certains domaines et sur certaines langues, la disponibilité de telles ressources peut être problématique. Il est alors nécessaire de mettre au point des méthodes d'extraction de termes à partir de textes spécialisés [Neifar and Ltaief, 2016]. C'est particulièrement le cas pour l'arabe standard moderne (MSA) où peu d'outils d'extraction de termes sont disponibles alors qu'il s'agit de la langue officielle de 26 pays et de plusieurs organismes internationaux comme l'Organisation Mondiale de la Santé (*OMS*).

Ainsi, dans cette thèse, nous nous sommes intéressés à la tâche d'extraction de terminologie en MSA. Nous avons exploré plusieurs stratégies d'extraction de termes simples ou complexes : adaptation d'un extracteur de termes existant au MSA, translittération de termes issus d'une langue source telle que l'anglais, transfert de termes extraits automatiquement de textes sources vers des textes arabes. Pour cela, nous avons dû constituer notre propre corpus parallèle pour mener à bien l'évaluation des approches proposées et identifier leur contribution à la constitution de ressources terminologiques en MSA. Cette évaluation suit un protocole que nous avons proposé. Celui-ci tient compte de la structure des résultats obtenus par chaque méthode en respectant deux conditions de base pour qu'un terme candidat soit considéré comme un terme du domaine : le terme doit être bien formé morpho-syntaxiquement et appartenir sémantiquement à notre domaine d'étude, ici le domaine médical et plus particulièrement la santé. Notre travail regroupe donc la collecte d'un corpus parallèle anglais-arabe et des contributions méthodologiques au niveau de l'extraction de termes en arabe.

La première contribution de ce travail de thèse consiste en la constitution d'un corpus parallèle de textes de spécialité. Nous avons choisi de nous focaliser sur des textes du domaine médical et notamment celui de la santé publique car il s'agit d'un domaine qui demande la disponibilité de l'information dans différentes langues. Nous avons choisi de traiter deux langues faisant partie des six langues officielles de l'Organisation Mondiale de la Santé *OMS* : l'arabe et l'anglais. Notre corpus est actuellement constitué de 102 textes parallèles anglais-arabe comprenant 79903 mots arabes désagglutinés et 60900 mots anglais. Afin d'analyser l'impact de la désagglutination des mots arabe lors de l'extraction de termes, nous avons également construit une version de notre corpus arabe où les clitiques sont séparés des mots auxquels ils sont associés. Ce corpus, dans sa version de base ou désagglutinée, a servi de source pour réaliser

nos expérimentations, avec des tailles variables selon son état d'avancement au moment des expériences et selon les données nécessaires aux tâches réalisées.

Nos principales contributions concernent la mise au point de plusieurs méthodes pour l'extraction de termes à partir de textes arabe. Dans un premier temps, nous avons adapté l'extracteur de termes YATEA afin de l'appliquer à des textes de spécialité en arabe standard moderne. Le processus d'extraction de termes candidats arabe a été défini, d'une part en s'appuyant sur une description des mécanismes de formation de la terminologie arabe, et d'autre part en prenant en compte les phénomènes d'agglutination, en particulier les proclitiques. Ainsi, l'analyse morphologique réalisée par MADA+TOKAN est exploitée pour mettre au point des patrons d'analyse syntaxique des termes candidats qui tiennent compte de ce phénomène. Des expériences ont été réalisées sur un corpus de textes médicaux arabes composé de 15532 mots. Celles-ci montrent une amélioration de la qualité des résultats lorsque les proclitiques sont pris en compte : le nombre de termes candidats extraits diminue et la précision des termes complexes maximaux augmente de 65,7 à 72,1% [Neifar et al., 2016a,b].

Comme la langue arabe se caractérise par l'assimilation de mots spécialisés issus d'autres langues, nous avons également proposé une méthode permettant d'extraire les termes arabes issus de la translittération de termes anglais en caractères arabes. Les termes simples anglais proposés par un extracteur de termes tel que YATEA sont tout d'abord translittérés grâce à une table de correspondance des caractères anglais en arabe que nous avons créée à partir d'une étude en corpus et de normes existantes pour la translittération de l'arabe vers l'anglais. Des traitements supplémentaires permettant de ternir compte du phénomène d'agglutination et de non voyellation des textes arabes sont ensuite appliqués pour améliorer la qualité des termes arabes proposés par cette méthode. Nous avons réalisé une évaluation de la méthode sur un corpus de textes médicaux parallèles anglais-arabe composé de 55208 mots anglais et 72204 mots arabes désagglutinés. L'analyse manuelle des résultats obtenus montre que 79,31% des termes anglais sont correctement translittérés et 67,15% des termes arabes extraits sont corrects [Neifar et al., 2018]. Outre l'extraction de termes simples arabes, cette méthode permet également d'enrichir une terminologie médicale bilingue.

Nous avons choisi de proposer une troisième méthode d'acquisition terminologique pour la langue arabe se basant sur la notion de transfert translingue. À partir d'un corpus de spécialité parallèle anglais-arabe, les termes arabes sont extraits grâce aux relations d'alignement pouvant exister entre les mots de notre corpus bilingue. Nous supposons alors que les expressions arabes qui correspondent aux termes candidats anglais extraits automatiquement représentent

des termes candidats arabes. Outre la réduction du coût de mise au point d'une méthode d'extraction de termes spécifiques à une langue, cette approche par transfert permet également de palier la carence de ressources ou d'outils de TAL pour l'arabe. Nous avons réalisé des évaluations sur un échantillon de 1000 couples de termes candidats tirés au hasard sur 6994 couples au total. Ces 1000 couples comportent 952 termes candidats arabes alignés avec 448 termes candidats anglais. L'étape d'évaluation nous permet d'extraire 358 termes candidats médicaux arabes (37,6%) bien formés morpho-syntaxiquement et faisant partie du domaine médical. Les termes arabes candidats rejetés sont répartis comme suit : i) 184 d'entre eux (soit 30,77% des termes candidats rejetés et 19,33% de la totalité) représentent des termes candidats qui sont bien formés morpho-syntaxiquement mais qui ne font pas partie du domaine médical, ii) 95 termes candidats arabes extraits font partie du domaine médical mais ne sont pas bien formés morpho-syntaxiquement (soit 15.87% des termes candidats rejetés et 9.98% de la totalité), iii) 315 termes candidats arabes extraits ne sont ni bien formés morpho-syntaxiquement, ni faisant partie du domaine médical (soit 52.67% des termes candidats rejetés et 33.09% de la totalité des termes candidats extraits). De plus, tout comme notre méthode d'extraction des termes translittérés, l'extraction des termes par transfert permet aussi d'enrichir une terminologie médicale bilingue. Sur l'ensemble de 1000 couples de l'échantillon, nous avons retenu 552 couples de termes candidats anglais-arabe ayant une correspondance complète, 129 couples ayant une correspondance partielle et 284 couples n'ayant pas de lien entre eux.

Les résultats obtenus par les différentes méthodes montrent que l'adaptation d'un extracteur de termes permet d'obtenir des termes bien formés mais avec un coût de mise au point important. Au contraire, l'extraction de termes par transfert permet de réduire considérablement l'effort de mise au point mais dépend grandement d'un bon alignement au niveau des mots. La translittération de termes anglais en arabe permet de compléter les listes de termes extraits en évitant les problèmes liées à un mauvais étiquetage morpho-syntaxique des mots empruntés.

6.2 Perspectives

Plusieurs perspectives de travail s'offrent à nous. Tout d'abord, nous souhaitons élargir notre corpus de textes anglais-arabe et lui ajouter des textes français parallèles. Nous envisageons par la suite de le mettre à disposition pour servir à d'autres travaux de recherche. Cet enrichissement de notre corpus permettrait, d'une part, d'améliorer les résultats obtenus lorsque plus de données sont nécessaires, et d'autre part, de nous assurer de la couverture des méthodes proposées ainsi que de la reproductibilité de nos résultats sur un autre couple de langues tout en conservant le

même domaine de spécialité. Nous envisageons également d'évaluer notre travail dans d'autres domaines de spécialité en fonction de la disponibilité de corpus. Dans ce contexte, l'application des méthodes proposées compléterait la caractérisation de leur couverture.

Enfin, les expériences menées jusqu'à présent ont permis d'identifier la précision des méthodes. Or il est également important de pouvoir évaluer leur rappel. Pour cela, nous devons, dans un premier temps, construire une référence en recensant les termes présents dans les textes de notre corpus de travail. Ce travail devra être réalisé sur des logiciels d'annotation offrant la manipulation des systèmes d'écriture de droite à gauche comme l'arabe. C'est, par exemple, le cas de Webanno[1] qui, contrairement à Brat[2], offre cette possibilité.

Sur le plan méthodologique, l'adaptation de YATEA pour le MSA peut être améliorée. Il s'agit notamment de mieux prendre en compte la voyellation et les phénomènes d'agglutination, notamment les enclitiques. L'amélioration peut également porter sur des traitements spécifiques corrigeant l'analyse morphologique fournie par MADA+TOKAN. Nous avons ainsi observé que lorsqu'un nom est dérivé d'un verbe qui désigne l'action associée (**ضمّادة** (*pansement*) /**ضمّد** (*panser*)), l'étiquetage morpho-syntaxique est souvent erroné, le nom étant considéré par MADA+TOKAN comme un verbe. Pour corriger ce type d'erreur, nous envisageons d'utiliser les marques morphologiques de cas ou le *masdar*. Une alternative pour améliorer les résultats de l'extraction de termes sera d'utiliser l'étiqueteur MADAMIRA. Outre l'augmentation de la couverture de l'extraction de termes, cette correction de l'étiquetage morpho-syntaxique permettrait d'améliorer l'analyse syntaxique des termes candidats proposés par YATEA. Nous tenons à mentionner que l'adaptation de YATEA pour le MSA sera disponible dans la prochaine version de YATEA.

En ce qui concerne la détection des termes médicaux anglais translittérés en caractères arabes, le processus d'alignement des textes au niveau des mots doit être amélioré. Nous proposons d'intégrer au processus d'alignement des caractéristiques morpho-syntaxiques des mots. Nous envisageons aussi d'extraire les termes translittérés à partir des bi-phrases plutôt qu'à partir des couples de termes candidats alignés anglais-arabe.

Par ailleurs, l'évaluation de cette approche sur d'autres textes de spécialité issus d'autres domaines mais aussi rédigés dans d'autres langues nous paraît particulièrement importante ici. En effet, l'origine des termes translittérés peut avoir une influence sur les résultats obtenus. Il est donc indispensable de mieux caractériser la contribution d'une telle approche d'extraction de termes en faisant varier ses contextes d'application.

1. `https://webanno.github.io/`
2. `http://brat.nlplab.org/`

Afin d'améliorer notre méthode d'extraction de termes par transfert, nous proposons de lui intégrer quelques patrons syntaxiques créés lors de l'adaptation de YATEA au MSA. Ceci permettra d'éviter l'extraction de termes candidats arabes mal formés faisant partie du domaine médical.

Enfin, pour compléter la réponse que nous avons apportée à la question de recherche soulevée dans cette thèse, nous envisageons de développer un extracteur de termes spécifique à l'arabe tout en tenant en compte des études et des analyses effectuées lors de l'application des différentes méthodes d'extraction présentées dans ce travail de thèse.

Annexe A

Table de correspondance des caractères anglais en arabe

Caractère anglais	Caractères arabes
a	ا ي ع ؤ ء ء
b	ب
c	ق س ك
d	ظ ض د
e	إ ا ي
f	ف
g	غ ج
h	ى ح ه
i	إ ئ ا ي
j	ج
k	ك
l	ل
m	م
n	ن
o	ا و
p	ب
q	ق
r	ر
s	ز ص س
t	ط ت
u	ئ و
v	ف
w	و
x	كس
y	ي
z	ز
ou	و
dh	ظ ض ذ
kh	خ
th	ت ث
ch	ك ش
sh	ش
gh	غ
ph	ف

Annexe B

Patrons syntaxiques arabes

(noun-m-s-n-c<=H> noun-m-s-n-d<=M>) 1 LEFT
(noun-m-s-n-c<=H> noun-f-s-n-d<=M>) 1 LEFT
(noun-m-s-n-c<=H> noun-m-d-n-d<=M>) 1 LEFT
(noun-m-s-n-c<=H> noun-f-d-n-d<=M>) 1 LEFT
(noun-m-s-n-c<=H> noun-m-p-n-d<=M>) 1 LEFT
(noun-m-s-n-c<=H> noun-f-p-n-d<=M>) 1 LEFT

(noun-f-s-n-c<=H> noun-m-s-n-d<=M>) 1 LEFT
(noun-f-s-n-c<=H> noun-f-s-n-d<=M>) 1 LEFT
(noun-f-s-n-c<=H> noun-m-d-n-d<=M>) 1 LEFT
(noun-f-s-n-c<=H> noun-f-d-n-d<=M>) 1 LEFT
(noun-f-s-n-c<=H> noun-m-p-n-d<=M>) 1 LEFT
(noun-f-s-n-c<=H> noun-f-p-n-d<=M>) 1 LEFT

(noun-m-d-n-c<=H> noun-m-s-n-d<=M>) 1 LEFT
(noun-m-d-n-c<=H> noun-f-s-n-d<=M>) 1 LEFT
(noun-m-d-n-c<=H> noun-m-d-n-d<=M>) 1 LEFT
(noun-m-d-n-c<=H> noun-f-d-n-d<=M>) 1 LEFT
(noun-m-d-n-c<=H> noun-m-p-n-d<=M>) 1 LEFT
(noun-m-d-n-c<=H> noun-f-p-n-d<=M>) 1 LEFT

(noun-f-d-n-c<=H> noun-m-s-n-d<=M>) 1 LEFT
(noun-f-d-n-c<=H> noun-f-s-n-d<=M>) 1 LEFT
(noun-f-d-n-c<=H> noun-m-d-n-d<=M>) 1 LEFT
(noun-f-d-n-c<=H> noun-f-d-n-d<=M>) 1 LEFT
(noun-f-d-n-c<=H> noun-m-p-n-d<=M>) 1 LEFT
(noun-f-d-n-c<=H> noun-f-p-n-d<=M>) 1 LEFT

(noun-m-p-n-c<=H> noun-m-s-n-d<=M>) 1 LEFT
(noun-m-p-n-c<=H> noun-f-s-n-d<=M>) 1 LEFT
(noun-m-p-n-c<=H> noun-m-d-n-d<=M>) 1 LEFT
(noun-m-p-n-c<=H> noun-f-d-n-d<=M>) 1 LEFT
(noun-m-p-n-c<=H> noun-m-p-n-d<=M>) 1 LEFT
(noun-m-p-n-c<=H> noun-f-p-n-d<=M>) 1 LEFT

(noun-f-p-n-c<=H> noun-m-s-n-d<=M>) 1 LEFT
(noun-f-p-n-c<=H> noun-f-s-n-d<=M>) 1 LEFT
(noun-f-p-n-c<=H> noun-m-d-n-d<=M>) 1 LEFT
(noun-f-p-n-c<=H> noun-f-d-n-d<=M>) 1 LEFT
(noun-f-p-n-c<=H> noun-m-p-n-d<=M>) 1 LEFT
(noun-f-p-n-c<=H> noun-f-p-n-d<=M>) 1 LEFT

(noun-m-s-n-c<=H> noun-m-s-n-c<=M>) 1 LEFT
(noun-m-s-n-c<=H> noun-f-s-n-c<=M>) 1 LEFT
(noun-m-s-n-c<=H> noun-m-d-n-c<=M>) 1 LEFT
(noun-m-s-n-c<=H> noun-f-d-n-c<=M>) 1 LEFT
(noun-m-s-n-c<=H> noun-m-p-n-c<=M>) 1 LEFT
(noun-m-s-n-c<=H> noun-f-p-n-c<=M>) 1 LEFT

(noun-f-s-n-c<=H> noun-m-s-n-c<=M>) 1 LEFT
(noun-f-s-n-c<=H> noun-f-s-n-c<=M>) 1 LEFT

(noun-f-s-n-c<=H> noun-m-d-n-c<=M>) 1 LEFT
(noun-f-s-n-c<=H> noun-f-d-n-c<=M>) 1 LEFT
(noun-f-s-n-c<=H> noun-m-p-n-c<=M>) 1 LEFT
(noun-f-s-n-c<=H> noun-f-p-n-c<=M>) 1 LEFT

(noun-m-d-n-c<=H> noun-m-s-n-c<=M>) 1 LEFT
(noun-m-d-n-c<=H> noun-f-s-n-c<=M>) 1 LEFT
(noun-m-d-n-c<=H> noun-m-d-n-c<=M>) 1 LEFT
(noun-m-d-n-c<=H> noun-f-d-n-c<=M>) 1 LEFT
(noun-m-d-n-c<=H> noun-m-p-n-c<=M>) 1 LEFT
(noun-m-d-n-c<=H> noun-f-p-n-c<=M>) 1 LEFT

(noun-f-d-n-c<=H> noun-m-s-n-c<=M>) 1 LEFT
(noun-f-d-n-c<=H> noun-f-s-n-c<=M>) 1 LEFT
(noun-f-d-n-c<=H> noun-m-d-n-c<=M>) 1 LEFT
(noun-f-d-n-c<=H> noun-f-d-n-c<=M>) 1 LEFT
(noun-f-d-n-c<=H> noun-m-p-n-c<=M>) 1 LEFT
(noun-f-d-n-c<=H> noun-f-p-n-c<=M>) 1 LEFT

(noun-m-p-n-c<=H> noun-m-s-n-c<=M>) 1 LEFT
(noun-m-p-n-c<=H> noun-f-s-n-c<=M>) 1 LEFT
(noun-m-p-n-c<=H> noun-m-d-n-c<=M>) 1 LEFT
(noun-m-p-n-c<=H> noun-f-d-n-c<=M>) 1 LEFT
(noun-m-p-n-c<=H> noun-m-p-n-c<=M>) 1 LEFT
(noun-m-p-n-c<=H> noun-f-p-n-c<=M>) 1 LEFT

(noun-m-s-n-c<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-m-s-n-c<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-m-s-n-c<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-m-s-n-c<=H> noun-f-d-g-d<=M>) 1 LEFT

```
( noun-m-s-n-c<=H> noun-m-p-g-d<=M> )    1    LEFT
( noun-m-s-n-c<=H> noun-f-p-g-d<=M> )    1    LEFT

( noun-f-s-n-c<=H> noun-m-s-g-d<=M> )    1    LEFT
( noun-f-s-n-c<=H> noun-f-s-g-d<=M> )    1    LEFT
( noun-f-s-n-c<=H> noun-m-d-g-d<=M> )    1    LEFT
( noun-f-s-n-c<=H> noun-f-d-g-d<=M> )    1    LEFT
( noun-f-s-n-c<=H> noun-m-p-g-d<=M> )    1    LEFT
( noun-f-s-n-c<=H> noun-f-p-g-d<=M> )    1    LEFT

( noun-m-d-n-c<=H> noun-m-s-g-d<=M> )    1    LEFT
( noun-m-d-n-c<=H> noun-f-s-g-d<=M> )    1    LEFT
( noun-m-d-n-c<=H> noun-m-d-g-d<=M> )    1    LEFT
( noun-m-d-n-c<=H> noun-f-d-g-d<=M> )    1    LEFT
( noun-m-d-n-c<=H> noun-m-p-g-d<=M> )    1    LEFT
( noun-m-d-n-c<=H> noun-f-p-g-d<=M> )    1    LEFT

( noun-f-d-n-c<=H> noun-m-s-g-d<=M> )    1    LEFT
( noun-f-d-n-c<=H> noun-f-s-g-d<=M> )    1    LEFT
( noun-f-d-n-c<=H> noun-m-d-g-d<=M> )    1    LEFT
( noun-f-d-n-c<=H> noun-f-d-g-d<=M> )    1    LEFT
( noun-f-d-n-c<=H> noun-m-p-g-d<=M> )    1    LEFT
( noun-f-d-n-c<=H> noun-f-p-g-d<=M> )    1    LEFT

( noun-m-p-n-c<=H> noun-m-s-g-d<=M> )    1    LEFT
( noun-m-p-n-c<=H> noun-f-s-g-d<=M> )    1    LEFT
( noun-m-p-n-c<=H> noun-m-d-g-d<=M> )    1    LEFT
( noun-m-p-n-c<=H> noun-f-d-g-d<=M> )    1    LEFT
( noun-m-p-n-c<=H> noun-m-p-g-d<=M> )    1    LEFT
( noun-m-p-n-c<=H> noun-f-p-g-d<=M> )    1    LEFT
```

(noun-f-p-n-c<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-f-p-n-c<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-f-p-n-c<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-f-p-n-c<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-f-p-n-c<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-f-p-n-c<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-m-s-g-d<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-m-s-g-d<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-m-s-g-d<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-m-s-g-d<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-m-s-g-d<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-m-s-g-d<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-f-s-g-d<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-f-s-g-d<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-f-s-g-d<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-f-s-g-d<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-f-s-g-d<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-f-s-g-d<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-m-d-g-d<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-m-d-g-d<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-m-d-g-d<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-m-d-g-d<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-m-d-g-d<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-m-d-g-d<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-f-d-g-d<=H> noun-m-s-g-d<=M>) 1 LEFT

(noun-f-d-g-d<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-f-d-g-d<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-f-d-g-d<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-f-d-g-d<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-f-d-g-d<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-m-p-g-d<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-m-p-g-d<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-m-p-g-d<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-m-p-g-d<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-m-p-g-d<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-m-p-g-d<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-f-p-g-d<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-f-p-g-d<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-f-p-g-d<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-f-p-g-d<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-f-p-g-d<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-f-p-g-d<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-m-s-g-i<=H> noun-m-s-g-i<=M>) 1 LEFT
(noun-m-s-g-i<=H> noun-f-s-g-i<=M>) 1 LEFT
(noun-m-s-g-i<=H> noun-m-d-g-i<=M>) 1 LEFT
(noun-m-s-g-i<=H> noun-f-d-g-i<=M>) 1 LEFT
(noun-m-s-g-i<=H> noun-m-p-g-i<=M>) 1 LEFT
(noun-m-s-g-i<=H> noun-f-p-g-i<=M>) 1 LEFT

(noun-f-s-g-i<=H> noun-m-s-g-i<=M>) 1 LEFT
(noun-f-s-g-i<=H> noun-f-s-g-i<=M>) 1 LEFT
(noun-f-s-g-i<=H> noun-m-d-g-i<=M>) 1 LEFT

(noun-f-s-g-i<=H> noun-f-d-g-i<=M>) 1 LEFT
(noun-f-s-g-i<=H> noun-m-p-g-i<=M>) 1 LEFT
(noun-f-s-g-i<=H> noun-f-p-g-i<=M>) 1 LEFT

(noun-m-d-g-i<=H> noun-m-s-g-i<=M>) 1 LEFT
(noun-m-d-g-i<=H> noun-f-s-g-i<=M>) 1 LEFT
(noun-m-d-g-i<=H> noun-m-d-g-i<=M>) 1 LEFT
(noun-m-d-g-i<=H> noun-f-d-g-i<=M>) 1 LEFT
(noun-m-d-g-i<=H> noun-m-p-g-i<=M>) 1 LEFT
(noun-m-d-g-i<=H> noun-f-p-g-i<=M>) 1 LEFT

(noun-f-d-g-i<=H> noun-m-s-g-i<=M>) 1 LEFT
(noun-f-d-g-i<=H> noun-f-s-g-i<=M>) 1 LEFT
(noun-f-d-g-i<=H> noun-m-d-g-i<=M>) 1 LEFT
(noun-f-d-g-i<=H> noun-f-d-g-i<=M>) 1 LEFT
(noun-f-d-g-i<=H> noun-m-p-g-i<=M>) 1 LEFT
(noun-f-d-g-i<=H> noun-f-p-g-i<=M>) 1 LEFT

(noun-m-p-g-i<=H> noun-m-s-g-i<=M>) 1 LEFT
(noun-m-p-g-i<=H> noun-f-s-g-i<=M>) 1 LEFT
(noun-m-p-g-i<=H> noun-m-d-g-i<=M>) 1 LEFT
(noun-m-p-g-i<=H> noun-f-d-g-i<=M>) 1 LEFT
(noun-m-p-g-i<=H> noun-m-p-g-i<=M>) 1 LEFT
(noun-m-p-g-i<=H> noun-f-p-g-i<=M>) 1 LEFT

(noun-f-p-g-i<=H> noun-m-s-g-i<=M>) 1 LEFT
(noun-f-p-g-i<=H> noun-f-s-g-i<=M>) 1 LEFT
(noun-f-p-g-i<=H> noun-m-d-g-i<=M>) 1 LEFT
(noun-f-p-g-i<=H> noun-f-d-g-i<=M>) 1 LEFT
(noun-f-p-g-i<=H> noun-m-p-g-i<=M>) 1 LEFT

(noun-f-p-g-i<=H> noun-f-p-g-i<=M>) 1 LEFT

(noun-m-s-g-d<=H> noun-m-s-g-i<=M>) 1 LEFT
(noun-m-s-g-d<=H> noun-f-s-g-i<=M>) 1 LEFT
(noun-m-s-g-d<=H> noun-m-d-g-i<=M>) 1 LEFT
(noun-m-s-g-d<=H> noun-f-d-g-i<=M>) 1 LEFT
(noun-m-s-g-d<=H> noun-m-p-g-i<=M>) 1 LEFT
(noun-m-s-g-d<=H> noun-f-p-g-i<=M>) 1 LEFT

(noun-f-s-g-d<=H> noun-m-s-g-i<=M>) 1 LEFT
(noun-f-s-g-d<=H> noun-f-s-g-i<=M>) 1 LEFT
(noun-f-s-g-d<=H> noun-m-d-g-i<=M>) 1 LEFT
(noun-f-s-g-d<=H> noun-f-d-g-i<=M>) 1 LEFT
(noun-f-s-g-d<=H> noun-m-p-g-i<=M>) 1 LEFT
(noun-f-s-g-d<=H> noun-f-p-g-i<=M>) 1 LEFT

(noun-m-d-g-d<=H> noun-m-s-g-i<=M>) 1 LEFT
(noun-m-d-g-d<=H> noun-f-s-g-i<=M>) 1 LEFT
(noun-m-d-g-d<=H> noun-m-d-g-i<=M>) 1 LEFT
(noun-m-d-g-d<=H> noun-f-d-g-i<=M>) 1 LEFT
(noun-m-d-g-d<=H> noun-m-p-g-i<=M>) 1 LEFT
(noun-m-d-g-d<=H> noun-f-p-g-i<=M>) 1 LEFT

(noun-f-d-g-d<=H> noun-m-s-g-i<=M>) 1 LEFT
(noun-f-d-g-d<=H> noun-f-s-g-i<=M>) 1 LEFT
(noun-f-d-g-d<=H> noun-m-d-g-i<=M>) 1 LEFT
(noun-f-d-g-d<=H> noun-f-d-g-i<=M>) 1 LEFT
(noun-f-d-g-d<=H> noun-m-p-g-i<=M>) 1 LEFT
(noun-f-d-g-d<=H> noun-f-p-g-i<=M>) 1 LEFT

(noun-m-p-g-d<=H> noun-m-s-g-i<=M>) 1 LEFT
(noun-m-p-g-d<=H> noun-f-s-g-i<=M>) 1 LEFT
(noun-m-p-g-d<=H> noun-m-d-g-i<=M>) 1 LEFT
(noun-m-p-g-d<=H> noun-f-d-g-i<=M>) 1 LEFT
(noun-m-p-g-d<=H> noun-m-p-g-i<=M>) 1 LEFT
(noun-m-p-g-d<=H> noun-f-p-g-i<=M>) 1 LEFT

(noun-f-p-g-d<=H> noun-m-s-g-i<=M>) 1 LEFT
(noun-f-p-g-d<=H> noun-f-s-g-i<=M>) 1 LEFT
(noun-f-p-g-d<=H> noun-m-d-g-i<=M>) 1 LEFT
(noun-f-p-g-d<=H> noun-f-d-g-i<=M>) 1 LEFT
(noun-f-p-g-d<=H> noun-m-p-g-i<=M>) 1 LEFT
(noun-f-p-g-d<=H> noun-f-p-g-i<=M>) 1 LEFT

(noun-m-s-g-c<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-f-s-g-c<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-m-d-g-c<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-f-s-g-d<=M>) 1 LEFT

(noun-m-d-g-c<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-f-d-g-c<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-m-p-g-c<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-f-p-g-c<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-m-s-g-c<=H> noun-m-s-g-c<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-f-s-g-c<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-m-d-g-c<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-f-d-g-c<=M>) 1 LEFT

(noun-m-s-g-c<=H> noun-m-p-g-c<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-f-p-g-c<=M>) 1 LEFT

(noun-f-s-g-c<=H> noun-m-s-g-c<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-f-s-g-c<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-m-d-g-c<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-f-d-g-c<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-m-p-g-c<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-f-p-g-c<=M>) 1 LEFT

(noun-m-d-g-c<=H> noun-m-s-g-c<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-f-s-g-c<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-m-d-g-c<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-f-d-g-c<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-m-p-g-c<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-f-p-g-c<=M>) 1 LEFT

(noun-f-d-g-c<=H> noun-m-s-g-c<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-f-s-g-c<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-m-d-g-c<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-f-d-g-c<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-m-p-g-c<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-f-p-g-c<=M>) 1 LEFT

(noun-m-p-g-c<=H> noun-m-s-g-c<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-f-s-g-c<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-m-d-g-c<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-f-d-g-c<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-m-p-g-c<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-f-p-g-c<=M>) 1 LEFT

(noun-f-p-g-c<=H> noun-m-s-g-c<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-f-s-g-c<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-m-d-g-c<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-f-d-g-c<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-m-p-g-c<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-f-p-g-c<=M>) 1 LEFT

(noun-m-s-g-c<=H> noun-m-s-g-i<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-f-s-g-i<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-m-d-g-i<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-f-d-g-i<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-m-p-g-i<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-f-p-g-i<=M>) 1 LEFT

(noun-f-s-g-c<=H> noun-m-s-g-i<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-f-s-g-i<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-m-d-g-i<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-f-d-g-i<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-m-p-g-i<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-f-p-g-i<=M>) 1 LEFT

(noun-m-d-g-c<=H> noun-m-s-g-i<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-f-s-g-i<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-m-d-g-i<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-f-d-g-i<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-m-p-g-i<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-f-p-g-i<=M>) 1 LEFT

(noun-f-d-g-c<=H> noun-m-s-g-i<=M>) 1 LEFT

(noun-f-d-g-c<=H> noun-f-s-g-i<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-m-d-g-i<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-f-d-g-i<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-m-p-g-i<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-f-p-g-i<=M>) 1 LEFT

(noun-m-p-g-c<=H> noun-m-s-g-i<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-f-s-g-i<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-m-d-g-i<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-f-d-g-i<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-m-p-g-i<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-f-p-g-i<=M>) 1 LEFT

(noun-f-p-g-c<=H> noun-m-s-g-i<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-f-s-g-i<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-m-d-g-i<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-f-d-g-i<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-m-p-g-i<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-f-p-g-i<=M>) 1 LEFT

(noun-m-s-g-c<=H> noun-m-s-u-d<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-f-s-u-d<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-m-d-u-d<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-f-d-u-d<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-m-p-u-d<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-f-p-u-d<=M>) 1 LEFT

(noun-f-s-g-c<=H> noun-m-s-u-d<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-f-s-u-d<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-m-d-u-d<=M>) 1 LEFT

(noun-f-s-g-c<=H> noun-f-d-u-d<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-m-p-u-d<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-f-p-u-d<=M>) 1 LEFT

(noun-m-d-g-c<=H> noun-m-s-u-d<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-f-s-u-d<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-m-d-u-d<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-f-d-u-d<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-m-p-u-d<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-f-p-u-d<=M>) 1 LEFT

(noun-f-d-g-c<=H> noun-m-s-u-d<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-f-s-u-d<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-m-d-u-d<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-f-d-u-d<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-m-p-u-d<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-f-p-u-d<=M>) 1 LEFT

(noun-m-p-g-c<=H> noun-m-s-u-d<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-f-s-u-d<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-m-d-u-d<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-f-d-u-d<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-m-p-u-d<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-f-p-u-d<=M>) 1 LEFT

(noun-f-p-g-c<=H> noun-m-s-u-d<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-f-s-u-d<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-m-d-u-d<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-f-d-u-d<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-m-p-u-d<=M>) 1 LEFT

(noun-f-p-g-c<=H> noun-f-p-u-d<=M>) 1 LEFT

(noun-m-s-g-c<=H> noun-m-s-n-d<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-f-s-n-d<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-m-d-n-d<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-f-d-n-d<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-m-p-n-d<=M>) 1 LEFT
(noun-m-s-g-c<=H> noun-f-p-n-d<=M>) 1 LEFT

(noun-f-s-g-c<=H> noun-m-s-n-d<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-f-s-n-d<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-m-d-n-d<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-f-d-n-d<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-m-p-n-d<=M>) 1 LEFT
(noun-f-s-g-c<=H> noun-f-p-n-d<=M>) 1 LEFT

(noun-m-d-g-c<=H> noun-m-s-n-d<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-f-s-n-d<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-m-d-n-d<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-f-d-n-d<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-m-p-n-d<=M>) 1 LEFT
(noun-m-d-g-c<=H> noun-f-p-n-d<=M>) 1 LEFT

(noun-f-d-g-c<=H> noun-m-s-n-d<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-f-s-n-d<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-m-d-n-d<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-f-d-n-d<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-m-p-n-d<=M>) 1 LEFT
(noun-f-d-g-c<=H> noun-f-p-n-d<=M>) 1 LEFT

(noun-m-p-g-c<=H> noun-m-s-n-d<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-f-s-n-d<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-m-d-n-d<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-f-d-n-d<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-m-p-n-d<=M>) 1 LEFT
(noun-m-p-g-c<=H> noun-f-p-n-d<=M>) 1 LEFT

(noun-f-p-g-c<=H> noun-m-s-n-d<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-f-s-n-d<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-m-d-n-d<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-f-d-n-d<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-m-p-n-d<=M>) 1 LEFT
(noun-f-p-g-c<=H> noun-f-p-n-d<=M>) 1 LEFT

(noun-m-s-g-i<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-m-s-g-i<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-m-s-g-i<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-m-s-g-i<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-m-s-g-i<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-m-s-g-i<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-f-s-g-i<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-f-s-g-i<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-f-s-g-i<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-f-s-g-i<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-f-s-g-i<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-f-s-g-i<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-m-d-g-i<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-m-d-g-i<=H> noun-f-s-g-d<=M>) 1 LEFT

(noun-m-d-g-i<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-m-d-g-i<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-m-d-g-i<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-m-d-g-i<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-f-d-g-i<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-f-d-g-i<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-f-d-g-i<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-f-d-g-i<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-f-d-g-i<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-f-d-g-i<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-m-p-g-i<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-m-p-g-i<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-m-p-g-i<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-m-p-g-i<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-m-p-g-i<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-m-p-g-i<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-f-p-g-i<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-f-p-g-i<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-f-p-g-i<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-f-p-g-i<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-f-p-g-i<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-f-p-g-i<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-m-s-a-i<=H> noun-m-s-g-c<=M>) 1 LEFT
(noun-m-s-a-i<=H> noun-f-s-g-c<=M>) 1 LEFT
(noun-m-s-a-i<=H> noun-m-d-g-c<=M>) 1 LEFT
(noun-m-s-a-i<=H> noun-f-d-g-c<=M>) 1 LEFT

```
( noun-m-s-a-i<=H> noun-m-p-g-c<=M> )    1    LEFT
( noun-m-s-a-i<=H> noun-f-p-g-c<=M> )    1    LEFT

( noun-f-s-a-i<=H> noun-m-s-g-c<=M> )    1    LEFT
( noun-f-s-a-i<=H> noun-f-s-g-c<=M> )    1    LEFT
( noun-f-s-a-i<=H> noun-m-d-g-c<=M> )    1    LEFT
( noun-f-s-a-i<=H> noun-f-d-g-c<=M> )    1    LEFT
( noun-f-s-a-i<=H> noun-m-p-g-c<=M> )    1    LEFT
( noun-f-s-a-i<=H> noun-f-p-g-c<=M> )    1    LEFT

( noun-m-d-a-i<=H> noun-m-s-g-c<=M> )    1    LEFT
( noun-m-d-a-i<=H> noun-f-s-g-c<=M> )    1    LEFT
( noun-m-d-a-i<=H> noun-m-d-g-c<=M> )    1    LEFT
( noun-m-d-a-i<=H> noun-f-d-g-c<=M> )    1    LEFT
( noun-m-d-a-i<=H> noun-m-p-g-c<=M> )    1    LEFT
( noun-m-d-a-i<=H> noun-f-p-g-c<=M> )    1    LEFT

( noun-f-d-a-i<=H> noun-m-s-g-c<=M> )    1    LEFT
( noun-f-d-a-i<=H> noun-f-s-g-c<=M> )    1    LEFT
( noun-f-d-a-i<=H> noun-m-d-g-c<=M> )    1    LEFT
( noun-f-d-a-i<=H> noun-f-d-g-c<=M> )    1    LEFT
( noun-f-d-a-i<=H> noun-m-p-g-c<=M> )    1    LEFT
( noun-f-d-a-i<=H> noun-f-p-g-c<=M> )    1    LEFT

( noun-m-p-a-i<=H> noun-m-s-g-c<=M> )    1    LEFT
( noun-m-p-a-i<=H> noun-f-s-g-c<=M> )    1    LEFT
( noun-m-p-a-i<=H> noun-m-d-g-c<=M> )    1    LEFT
( noun-m-p-a-i<=H> noun-f-d-g-c<=M> )    1    LEFT
( noun-m-p-a-i<=H> noun-m-p-g-c<=M> )    1    LEFT
( noun-m-p-a-i<=H> noun-f-p-g-c<=M> )    1    LEFT
```

(noun-f-p-a-i<=H> noun-m-s-g-c<=M>) 1 LEFT
(noun-f-p-a-i<=H> noun-f-s-g-c<=M>) 1 LEFT
(noun-f-p-a-i<=H> noun-m-d-g-c<=M>) 1 LEFT
(noun-f-p-a-i<=H> noun-f-d-g-c<=M>) 1 LEFT
(noun-f-p-a-i<=H> noun-m-p-g-c<=M>) 1 LEFT
(noun-f-p-a-i<=H> noun-f-p-g-c<=M>) 1 LEFT

(noun-m-s-a-i<=H> noun-m-s-a-d<=M>) 1 LEFT
(noun-m-s-a-i<=H> noun-f-s-a-d<=M>) 1 LEFT
(noun-m-s-a-i<=H> noun-m-d-a-d<=M>) 1 LEFT
(noun-m-s-a-i<=H> noun-f-d-a-d<=M>) 1 LEFT
(noun-m-s-a-i<=H> noun-m-p-a-d<=M>) 1 LEFT
(noun-m-s-a-i<=H> noun-f-p-a-d<=M>) 1 LEFT

(noun-f-s-a-i<=H> noun-m-s-a-d<=M>) 1 LEFT
(noun-f-s-a-i<=H> noun-f-s-a-d<=M>) 1 LEFT
(noun-f-s-a-i<=H> noun-m-d-a-d<=M>) 1 LEFT
(noun-f-s-a-i<=H> noun-f-d-a-d<=M>) 1 LEFT
(noun-f-s-a-i<=H> noun-m-p-a-d<=M>) 1 LEFT
(noun-f-s-a-i<=H> noun-f-p-a-d<=M>) 1 LEFT

(noun-m-d-a-i<=H> noun-m-s-a-d<=M>) 1 LEFT
(noun-m-d-a-i<=H> noun-f-s-a-d<=M>) 1 LEFT
(noun-m-d-a-i<=H> noun-m-d-a-d<=M>) 1 LEFT
(noun-m-d-a-i<=H> noun-f-d-a-d<=M>) 1 LEFT
(noun-m-d-a-i<=H> noun-m-p-a-d<=M>) 1 LEFT
(noun-m-d-a-i<=H> noun-f-p-a-d<=M>) 1 LEFT

(noun-f-d-a-i<=H> noun-m-s-a-d<=M>) 1 LEFT

```
( noun-f-d-a-i<=H> noun-f-s-a-d<=M> )      1   LEFT
( noun-f-d-a-i<=H> noun-m-d-a-d<=M> )      1   LEFT
( noun-f-d-a-i<=H> noun-f-d-a-d<=M> )      1   LEFT
( noun-f-d-a-i<=H> noun-m-p-a-d<=M> )      1   LEFT
( noun-f-d-a-i<=H> noun-f-p-a-d<=M> )      1   LEFT

( noun-m-p-a-i<=H> noun-m-s-a-d<=M> )      1   LEFT
( noun-m-p-a-i<=H> noun-f-s-a-d<=M> )      1   LEFT
( noun-m-p-a-i<=H> noun-m-d-a-d<=M> )      1   LEFT
( noun-m-p-a-i<=H> noun-f-d-a-d<=M> )      1   LEFT
( noun-m-p-a-i<=H> noun-m-p-a-d<=M> )      1   LEFT
( noun-m-p-a-i<=H> noun-f-p-a-d<=M> )      1   LEFT

( noun-f-p-a-i<=H> noun-m-s-a-d<=M> )      1   LEFT
( noun-f-p-a-i<=H> noun-f-s-a-d<=M> )      1   LEFT
( noun-f-p-a-i<=H> noun-m-d-a-d<=M> )      1   LEFT
( noun-f-p-a-i<=H> noun-f-d-a-d<=M> )      1   LEFT
( noun-f-p-a-i<=H> noun-m-p-a-d<=M> )      1   LEFT
( noun-f-p-a-i<=H> noun-f-p-a-d<=M> )      1   LEFT

( noun-m-s-a-i<=H> noun-m-s-g-d<=M> )      1   LEFT
( noun-m-s-a-i<=H> noun-f-s-g-d<=M> )      1   LEFT
( noun-m-s-a-i<=H> noun-m-d-g-d<=M> )      1   LEFT
( noun-m-s-a-i<=H> noun-f-d-g-d<=M> )      1   LEFT
( noun-m-s-a-i<=H> noun-m-p-g-d<=M> )      1   LEFT
( noun-m-s-a-i<=H> noun-f-p-g-d<=M> )      1   LEFT

( noun-f-s-a-i<=H> noun-m-s-g-d<=M> )      1   LEFT
( noun-f-s-a-i<=H> noun-f-s-g-d<=M> )      1   LEFT
( noun-f-s-a-i<=H> noun-m-d-g-d<=M> )      1   LEFT
```

(noun-f-s-a-i<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-f-s-a-i<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-f-s-a-i<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-m-d-a-i<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-m-d-a-i<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-m-d-a-i<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-m-d-a-i<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-m-d-a-i<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-m-d-a-i<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-f-d-a-i<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-f-d-a-i<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-f-d-a-i<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-f-d-a-i<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-f-d-a-i<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-f-d-a-i<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-m-p-a-i<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-m-p-a-i<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-m-p-a-i<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-m-p-a-i<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-m-p-a-i<=H> noun-m-p-g-d<=M>) 1 LEFT
(noun-m-p-a-i<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-f-p-a-i<=H> noun-m-s-g-d<=M>) 1 LEFT
(noun-f-p-a-i<=H> noun-f-s-g-d<=M>) 1 LEFT
(noun-f-p-a-i<=H> noun-m-d-g-d<=M>) 1 LEFT
(noun-f-p-a-i<=H> noun-f-d-g-d<=M>) 1 LEFT
(noun-f-p-a-i<=H> noun-m-p-g-d<=M>) 1 LEFT

(noun-f-p-a-i<=H> noun-f-p-g-d<=M>) 1 LEFT

(noun-m-s-g-i<=H> noun-m-s-g-c<=M>) 1 LEFT
(noun-m-s-g-i<=H> noun-f-s-g-c<=M>) 1 LEFT
(noun-m-s-g-i<=H> noun-m-d-g-c<=M>) 1 LEFT
(noun-m-s-g-i<=H> noun-f-d-g-c<=M>) 1 LEFT
(noun-m-s-g-i<=H> noun-m-p-g-c<=M>) 1 LEFT
(noun-m-s-g-i<=H> noun-f-p-g-c<=M>) 1 LEFT

(noun-f-s-g-i<=H> noun-m-s-g-c<=M>) 1 LEFT
(noun-f-s-g-i<=H> noun-f-s-g-c<=M>) 1 LEFT
(noun-f-s-g-i<=H> noun-m-d-g-c<=M>) 1 LEFT
(noun-f-s-g-i<=H> noun-f-d-g-c<=M>) 1 LEFT
(noun-f-s-g-i<=H> noun-m-p-g-c<=M>) 1 LEFT
(noun-f-s-g-i<=H> noun-f-p-g-c<=M>) 1 LEFT

(noun-m-d-g-i<=H> noun-m-s-g-c<=M>) 1 LEFT
(noun-m-d-g-i<=H> noun-f-s-g-c<=M>) 1 LEFT
(noun-m-d-g-i<=H> noun-m-d-g-c<=M>) 1 LEFT
(noun-m-d-g-i<=H> noun-f-d-g-c<=M>) 1 LEFT
(noun-m-d-g-i<=H> noun-m-p-g-c<=M>) 1 LEFT
(noun-m-d-g-i<=H> noun-f-p-g-c<=M>) 1 LEFT

(noun-f-d-g-i<=H> noun-m-s-g-c<=M>) 1 LEFT
(noun-f-d-g-i<=H> noun-f-s-g-c<=M>) 1 LEFT
(noun-f-d-g-i<=H> noun-m-d-g-c<=M>) 1 LEFT
(noun-f-d-g-i<=H> noun-f-d-g-c<=M>) 1 LEFT
(noun-f-d-g-i<=H> noun-m-p-g-c<=M>) 1 LEFT
(noun-f-d-g-i<=H> noun-f-p-g-c<=M>) 1 LEFT

(noun-m-p-g-i<=H> noun-m-s-g-c<=M>) 1 LEFT
(noun-m-p-g-i<=H> noun-f-s-g-c<=M>) 1 LEFT
(noun-m-p-g-i<=H> noun-m-d-g-c<=M>) 1 LEFT
(noun-m-p-g-i<=H> noun-f-d-g-c<=M>) 1 LEFT
(noun-m-p-g-i<=H> noun-m-p-g-c<=M>) 1 LEFT
(noun-m-p-g-i<=H> noun-f-p-g-c<=M>) 1 LEFT

(noun-f-p-g-i<=H> noun-m-s-g-c<=M>) 1 LEFT
(noun-f-p-g-i<=H> noun-f-s-g-c<=M>) 1 LEFT
(noun-f-p-g-i<=H> noun-m-d-g-c<=M>) 1 LEFT
(noun-f-p-g-i<=H> noun-f-d-g-c<=M>) 1 LEFT
(noun-f-p-g-i<=H> noun-m-p-g-c<=M>) 1 LEFT
(noun-f-p-g-i<=H> noun-f-p-g-c<=M>) 1 LEFT

(noun-m-s-n-d<=H> adj-m-s-a-d<=M>) 1 LEFT
(noun-f-s-n-d<=H> adj-f-s-a-d<=M>) 1 LEFT
(noun-m-d-n-d<=H> adj-m-d-a-d<=M>) 1 LEFT
(noun-f-d-n-d<=H> adj-f-d-a-d<=M>) 1 LEFT
(noun-m-p-n-d<=H> adj-m-p-a-d<=M>) 1 LEFT
(noun-f-p-n-d<=H> adj-f-p-a-d<=M>) 1 LEFT
(noun-f-p-n-d<=H> adj-f-s-a-d<=M>) 1 LEFT

(noun-m-s-n-i<=H> adj-m-s-a-i<=M>) 1 LEFT
(noun-f-s-n-i<=H> adj-f-s-a-i<=M>) 1 LEFT
(noun-m-d-n-i<=H> adj-m-d-a-i<=M>) 1 LEFT
(noun-f-d-n-i<=H> adj-f-d-a-i<=M>) 1 LEFT
(noun-m-p-n-i<=H> adj-m-p-a-i<=M>) 1 LEFT
(noun-f-p-n-i<=H> adj-f-p-a-i<=M>) 1 LEFT
(noun-f-p-n-i<=H> adj-f-s-a-i<=M>) 1 LEFT

(noun-m-s-g-d<=H> adj-m-s-u-d<=M>) 1 LEFT
(noun-f-s-g-d<=H> adj-f-s-u-d<=M>) 1 LEFT
(noun-m-d-g-d<=H> adj-m-d-u-d<=M>) 1 LEFT
(noun-f-d-g-d<=H> adj-f-d-u-d<=M>) 1 LEFT
(noun-m-p-g-d<=H> adj-m-p-u-d<=M>) 1 LEFT
(noun-f-p-g-d<=H> adj-f-p-u-d<=M>) 1 LEFT
(noun-f-p-g-d<=H> adj-f-s-u-d<=M>) 1 LEFT

(noun-m-s-u-d<=H> adj-m-s-u-d<=M>) 1 LEFT
(noun-f-s-u-d<=H> adj-f-s-u-d<=M>) 1 LEFT
(noun-m-d-u-d<=H> adj-m-d-u-d<=M>) 1 LEFT
(noun-f-d-u-d<=H> adj-f-d-u-d<=M>) 1 LEFT
(noun-m-p-u-d<=H> adj-m-p-u-d<=M>) 1 LEFT
(noun-f-p-u-d<=H> adj-f-p-u-d<=M>) 1 LEFT
(noun-f-p-u-d<=H> adj-f-s-u-d<=M>) 1 LEFT

(noun-m-s-n-d<=H> adj-m-s-a-i<=M>) 1 LEFT
(noun-f-s-n-d<=H> adj-f-s-a-i<=M>) 1 LEFT
(noun-m-d-n-d<=H> adj-m-d-a-i<=M>) 1 LEFT
(noun-f-d-n-d<=H> adj-f-d-a-i<=M>) 1 LEFT
(noun-m-p-n-d<=H> adj-m-p-a-i<=M>) 1 LEFT
(noun-f-p-n-d<=H> adj-f-p-a-i<=M>) 1 LEFT
(noun-f-p-n-d<=H> adj-f-s-a-i<=M>) 1 LEFT

(noun-m-s-n-d<=H> adj-m-s-g-d<=M>) 1 LEFT
(noun-f-s-n-d<=H> adj-f-s-g-d<=M>) 1 LEFT
(noun-m-d-n-d<=H> adj-m-d-g-d<=M>) 1 LEFT
(noun-f-d-n-d<=H> adj-f-d-g-d<=M>) 1 LEFT
(noun-m-p-n-d<=H> adj-m-p-g-d<=M>) 1 LEFT
(noun-f-p-n-d<=H> adj-f-p-g-d<=M>) 1 LEFT

(noun-f-p-n-d<=H> adj-f-s-g-d<=M>) 1 LEFT

(noun-m-s-n-d<=H> adj-m-s-n-d<=M>) 1 LEFT
(noun-f-s-n-d<=H> adj-f-s-n-d<=M>) 1 LEFT
(noun-m-d-n-d<=H> adj-m-d-n-d<=M>) 1 LEFT
(noun-f-d-n-d<=H> adj-f-d-n-d<=M>) 1 LEFT
(noun-m-p-n-d<=H> adj-m-p-n-d<=M>) 1 LEFT
(noun-f-p-n-d<=H> adj-f-p-n-d<=M>) 1 LEFT
(noun-f-p-n-d<=H> adj-f-s-n-d<=M>) 1 LEFT

(noun-m-s-g-d<=H> adj-m-s-g-d<=M>) 1 LEFT
(noun-f-s-g-d<=H> adj-f-s-g-d<=M>) 1 LEFT
(noun-m-d-g-d<=H> adj-m-d-g-d<=M>) 1 LEFT
(noun-f-d-g-d<=H> adj-f-d-g-d<=M>) 1 LEFT
(noun-m-p-g-d<=H> adj-m-p-g-d<=M>) 1 LEFT
(noun-f-p-g-d<=H> adj-f-p-g-d<=M>) 1 LEFT
(noun-f-p-g-d<=H> adj-f-s-g-d<=M>) 1 LEFT

(noun-m-s-g-c<=H> adj-m-s-g-c<=M>) 1 LEFT
(noun-f-s-g-c<=H> adj-f-s-g-c<=M>) 1 LEFT
(noun-m-d-g-c<=H> adj-m-d-g-c<=M>) 1 LEFT
(noun-f-d-g-c<=H> adj-f-d-g-c<=M>) 1 LEFT
(noun-m-p-g-c<=H> adj-m-p-g-c<=M>) 1 LEFT
(noun-f-p-g-c<=H> adj-f-p-g-c<=M>) 1 LEFT
(noun-f-p-g-c<=H> adj-f-s-g-c<=M>) 1 LEFT

(noun-m-s-g-c<=H> adj-m-s-g-i<=M>) 1 LEFT
(noun-f-s-g-c<=H> adj-f-s-g-i<=M>) 1 LEFT
(noun-m-d-g-c<=H> adj-m-d-g-i<=M>) 1 LEFT
(noun-f-d-g-c<=H> adj-f-d-g-i<=M>) 1 LEFT

(noun-m-p-g-c<=H> adj-m-p-g-i<=M>) 1 LEFT
(noun-f-p-g-c<=H> adj-f-p-g-i<=M>) 1 LEFT
(noun-f-p-g-c<=H> adj-f-s-g-i<=M>) 1 LEFT

(noun-m-s-a-d<=H> adj-m-s-g-d<=M>) 1 LEFT
(noun-f-s-a-d<=H> adj-f-s-g-d<=M>) 1 LEFT
(noun-m-d-a-d<=H> adj-m-d-g-d<=M>) 1 LEFT
(noun-f-d-a-d<=H> adj-f-d-g-d<=M>) 1 LEFT
(noun-m-p-a-d<=H> adj-m-p-g-d<=M>) 1 LEFT
(noun-f-p-a-d<=H> adj-f-p-g-d<=M>) 1 LEFT
(noun-f-p-a-d<=H> adj-f-s-g-d<=M>) 1 LEFT

(noun-m-s-a-d<=H> adj-m-s-a-d<=M>) 1 LEFT
(noun-f-s-a-d<=H> adj-f-s-a-d<=M>) 1 LEFT
(noun-m-d-a-d<=H> adj-m-d-a-d<=M>) 1 LEFT
(noun-f-d-a-d<=H> adj-f-d-a-d<=M>) 1 LEFT
(noun-m-p-a-d<=H> adj-m-p-a-d<=M>) 1 LEFT
(noun-f-p-a-d<=H> adj-f-p-a-d<=M>) 1 LEFT
(noun-f-p-a-d<=H> adj-f-s-a-d<=M>) 1 LEFT

(noun-m-s-g-i<=H> adj-m-s-g-i<=M>) 1 LEFT
(noun-f-s-g-i<=H> adj-f-s-g-i<=M>) 1 LEFT
(noun-m-d-g-i<=H> adj-m-d-g-i<=M>) 1 LEFT
(noun-f-d-g-i<=H> adj-f-d-g-i<=M>) 1 LEFT
(noun-m-p-g-i<=H> adj-m-p-g-i<=M>) 1 LEFT
(noun-f-p-g-i<=H> adj-f-p-g-i<=M>) 1 LEFT
(noun-f-p-g-i<=H> adj-f-s-g-i<=M>) 1 LEFT
(noun-m-s-g-i<=H> adj-f-s-g-i<=M>) 1 LEFT

(noun-m-s-g-d<=H> adj-m-s-u-d<=M>) 1 LEFT

(adj_comp-m-s-a-c<=M> noun-m-s-a-d<=H>) 1 RIGHT
(adj_comp-m-s-a-c<=M> noun-m-d-a-d<=H>) 1 RIGHT
(adj_comp-m-s-a-c<=M> noun-m-p-a-d<=H>) 1 RIGHT
(adj_comp-m-s-a-c<=M> noun-f-s-a-d<=H>) 1 RIGHT
(adj_comp-m-s-a-c<=M> noun-f-d-a-d<=H>) 1 RIGHT
(adj_comp-m-s-a-c<=M> noun-f-p-a-d<=H>) 1 RIGHT

(noun-m-s-n-c<=H> (noun-m-s-na-na<=C1> أو noun-m-s-na-na<=C2>)<=M>) 1 LEFT
(noun-m-s-n-c<=H> (noun-m-s-na-na<=C1> أو noun-m-s-na-na<=C2>)<=M>) 1 LEFT
(noun-f-p-g-c<=H> (noun-m-s-g-d<=C1> أو noun-m-s-g-d<=C2>)<=M>) 1 LEFT
(noun-f-s-a-i<=H> (noun-m-s-g-d<=C1> أو noun-m-s-g-d<=C2>)<=M>) 1 LEFT
(noun-m-s-g-d<=H> (noun-m-s-g-d<=C1> أو noun-m-s-g-d<=C2>)<=M>) 1 LEFT
(noun-f-s-a-i<=H> (noun-m-d-a-d<=C1> أو noun-m-d-g-d<=C2>)<=M>) 1 LEFT
(noun-f-s-a-i<=H> (noun-m-d-a-d<=C1> أو noun-m-d-g-d<=C2>)<=M>) 1 LEFT
(noun-f-s-a-i<=H> (noun-m-s-a-d<=C1> أو noun-m-s-g-d<=C2>)<=M>) 1 LEFT
(noun-f-s-a-i<=H> (noun-m-p-a-d<=C1> أو noun-m-p-g-d<=C2>)<=M>) 1 LEFT
(noun-f-s-a-i<=H> (noun-f-d-a-d<=C1> أو noun-f-d-g-d<=C2>)<=M>) 1 LEFT
(noun-f-s-a-i<=H> (noun-f-s-a-d<=C1> أو noun-f-s-g-d<=C2>)<=M>) 1 LEFT
(noun-f-s-a-i<=H> (noun-f-p-a-d<=C1> أو noun-f-p-g-d<=C2>)<=M>) 1 LEFT

(noun-f-s-n-d<=H> (adj-f-s-g-i<=M> noun-m-s-g-d<=H>)<=M>) 1 LEFT
(noun-m-s-g-d<=C1> أو) noun-f-s-g-c<=H> noun-m-d-g-d<=M>)<=C2>) 1 LEFT
(noun-m-s-n-i<=C1> أو) noun-m-s-g-i<=H> adj-m-s-g-i<=M>)<=C2>) 1 LEFT
(noun-m-s-n-d<=H> من) noun-f-s-g-c<=H> noun-m-s-g-d<=M>)<=M>) 1 LEFT
(noun-f-s-g-d<=H> من noun-f-s-g-d<=M>) 1 LEFT
(noun-m-s-g-i<=H> في noun-f-s-g-c<=M>) 1 LEFT
(noun-m-s-g-d<=H> في noun-m-s-g-d<=M>) 1 LEFT
(noun-f-s-a-i<=H> في noun-m-s-g-d<=M>) 1 LEFT

((noun-f-s-g-i<=M> adj-f-s-g-i<=H>)<=H> adj-f-s-g-i<=M>)	1	LEFT
(noun-m-s-g-c<=H> (noun-m-s-g-c<=H> noun-f-d-g-d<=M>)<=M>)	1	LEFT
(noun-m-s-g-d<=C1> أو noun-m-s-g-d<=C2>)	1	LEFT
(noun-f-s-g-d<=C1> أو noun-f-s-g-d<=C2>)	1	LEFT
(noun-m-d-g-d<=C1> أو noun-m-d-g-d<=C2>)	1	LEFT
(noun-f-d-g-d<=C1> أو noun-f-d-g-d<=C2>)	1	LEFT
(noun-m-p-g-d<=C1> أو noun-m-p-g-d<=C2>)	1	LEFT
(noun-f-p-g-d<=C1> أو noun-f-p-g-d<=C2>)	1	LEFT
(noun-m-d-a-d<=C1> أو noun-m-d-g-d<=C2>)	1	LEFT
(noun-m-s-g-c<=C1> أو noun-m-s-g-c<=C2>)	1	LEFT
(noun-m-s-g-c<=C1> أو noun-m-s-g-d<=C2>)	1	LEFT
(noun-m-s-g-d<=C1> أو noun-f-s-g-c<=C2>)	1	LEFT
(noun-m-s-n-c<=C1> أو noun-m-s-g-d<=C2>)	1	LEFT
(noun-m-s-n-c<=C1> أَوْ noun-m-s-g-c<=C2>)	1	LEFT
(noun-m-s-n-d<=C1> أو noun-m-s-g-c<=C2>)	1	LEFT
(noun-m-s-n-d<=C1> أو noun-m-s-g-d<=C2>)	1	LEFT
(noun-f-p-n-d<=C1> أو noun-f-p-g-c<=C2>)	1	LEFT
(noun-m-s-g-i<=C1> أو noun-m-s-g-c<=C2>)	1	LEFT
(noun-m-s-a-d<=H> على noun-f-s-g-d<=M>)	1	LEFT
(noun-m-s-g-d<=H> على noun-m-s-g-d<=M>)	1	LEFT
(noun-f-s-g-d<=H> عن noun-m-s-g-d<=M>)	1	LEFT
(noun-m-d-g-c<=C1> وَ noun-m-s-n-c<=C2>)	1	LEFT
(noun-m-s-a-c<=C1> و noun-m-s-n-c<=C2>)	1	LEFT
(noun-m-s-g-d<=C1> و noun-m-s-n-d<=C2>)	1	LEFT
(noun-m-p-a-c<=C1> و noun-m-s-n-c<=C2>)	1	LEFT
(noun-m-s-g-d<=C1> و noun-f-s-g-d<=C2>)	1	LEFT
(noun-m-d-n-d<=C1> و noun-m-d-n-d<=C2>)	1	LEFT
(noun-f-p-g-d<=C1> و noun-m-s-g-d<=C2>)	1	LEFT

(noun-m-s-g-d<=C1> و noun-f-s-n-d<=C2>)	1	LEFT
(noun-m-d-g-c<=C1> و noun-m-s-n-c<=C2>)	1	LEFT
(noun-m-s-g-d<=H> (noun-m-s-g-d<=C1> و noun-m-s-n-d<=C2>)<=M>)	1	LEFT
(noun-m-s-n-c<=H> (noun-f-s-g-d<=C1> و noun-m-s-g-d<=C2>)<=M>)	1	LEFT
(noun-m-s-g-c<=H> (noun-m-s-g-d<=C1> و noun-m-s-n-d<=C2>)<=M>)	1	LEFT
((noun-m-s-n-i<=C1> و noun-m-s-n-c<=C2>)<=H> noun-m-s-g-d)<=M>)	1	LEFT
(noun-m-s-a-c<=H> noun-m-s-a-c<=M>)	1	LEFT

Bibliographie

Ali Mashaan Abed, Sabrine Tiun, and Mohammed Albared. Arabic term extraction using combined approach on Islamic document. *Journal of Theoretical & Applied Information Technology*, 58(3), 2013.

Carine Abi Ghanem-Chadarevian. Socioterminologie et interactions langagières en arabe. *Repères DoRiF - Le terme : un produit social ?*, 10, April 2016.

Yaser Al-Onaizan and Kevin Knight. Translating named entities using monolingual and bilingual resources. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 400-408, 2002. doi : 10.3115/1073083.1073150. URL `https://doi.org/10.3115/1073083.1073150`.

Latifa Al-Sulaiti and Eric Atwell. The design of a corpus of Contemporary Arabic. *International Journal of Corpus Linguistics*, 11(1) :1-36, 2006.

Fahad Albogamy and Allan Ramsay. POS tagging for Arabic tweets. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 1-8, September 2015. URL `http://www.aclweb.org/anthology/R15-1001`.

Khalid AlKhatib and Amer Badarneh. Automatic extraction of Arabic multi-word terms. In *IMCSIT*, pages 411-418, 2010.

Sophie Aubin and Thierry Hamon. Improving term extraction with terminological resources. In Tapio Salakoski, Filip Ginter, Sampo Pyysalo, and Tapio Pahikkala, editors, *5th International Conference on NLP (FinTAL 2006)*, number 4139 in LNAI, pages 380-387. Springer, August 2006.

Raja Ayed, Ibrahim Bounhas, Bilel Elayeb, Fabrice Evrard, and Narjès Bellamine Ben Saoud. A possibilistic approach for the automatic morphological disambiguation of Arabic texts.

In *13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2012, Kyoto, Japan, August 8-10, 2012*, pages 187-194, 2012. doi : 10.1109/SNPD.2012.21. URL `https://doi.org/10.1109/SNPD.2012.21`.

Emile Benveniste. *Formes nouvelles de la composition nominale*. Klincksieck, 1966. URL `https://books.google.tn/books?id=OVVgOQAACAAJ`.

Francesca Bonin, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. A contrastive approach to multi-word term extraction from domain corpora. *Proc. of the 7th Int'l Conf. on Language Resources and Evaluation (LREC 2010)*, page 3222–3229, May 2010.

Dhouha Bouamor. *Using parallel and comparable corpora for multilingual linguistic resources extraction*. Theses, Université Paris Sud - Paris XI, February 2014.

Siham Boulaknadel, Beatrice Daille, and Driss Aboutajdine. A multi-word term extraction program for Arabic language. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the LREC'08*, may 2008. ISBN 2-9517408-4-0.

Ibrahim Bounhas and Yahya Slimani. A hybrid approach for Arabic multi-word term extraction. In *Natural Language Processing and Knowledge Engineering, NLP-KE 2009. International Conference on*, pages 1-8. IEEE, IEEE, 2009.

Ibrahim Bounhas, Wiem Lahbib, and Bilel Elayeb. Arabic domain terminology extraction : A literature review - (short paper). In *OTM 2014 Conferences - Confederated International Conferences : CoopIS, and ODBASE 2014*, pages 792-799, Amantea, Italy, October 2014.

Didier Bourigault. An endogeneous corpus-based method for structural noun phrase disambiguation. In *Proceedings of the EACL'93*, pages 81-86, Utrecht, The Netherlands, April 1993.

Didier Bourigault. *LEXTER un Logiciel d'EXtraction de TERminologie. Application à l'extraction des connaissances à partir de textes*. Thèse de doctorat (spécialité mathématiques, informatique appliquée aux sciences de l'homme, École des Hautes Études en Sciences Sociales (EHESS), Paris, France, 1994.

Didier Bourigault, Nathalie Aussenac-Gilles, and Jean Charlet. Construction de ressources terminologiques ou ontologiques à partir de textes : Un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle*, 18 :87-110, 02 2004.

Tim Buckwalter. Buckwalter arabic morphological analyzer version 1.0. *Linguistic Data Consortium (LDC)*, 01 2002.

M. Teresa Cabré, R. Estopà, and J. Vivaldi. Automatic term detection : a review of current systems. In *Recent Advances in Computational Terminology*. John Benjamins, 2001.

Maria Teresa Cabré Castellví. Theories of terminology. their description, prescription and explanation. *Terminology*, 9(2) :163-199, 2003. ISSN 0929-9971. doi : 10.1075/term.9.2.03cab. URL `http://dx.doi.org/10.1075/term.9.2.03cab`.

Yun-Chuang Chiao. *Bilingual lexicon extraction from comparable medical texts : application for cross-language information retrieval*. Theses, Université Pierre et Marie Curie - Paris VI, June 2004. URL `https://tel.archives-ouvertes.fr/tel-00007704`.

Kevin Bretonnel Cohen and Dina Demner-Fushman. *Biomedical Natural Language Processing*. John Benjamins publishing company, 2013.

Merley Conrado, Thiago Pardo, and Solange Rezende. A machine learning approach to automatic term extraction using a rich feature set. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*, pages 16-23, Atlanta, Georgia, June 2013. URL `http://www.aclweb.org/anthology/N13-2003`.

Ido Dagan, Alon Itai, and Ulrike Schwall. Two languages are more informative than one. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, ACL '91, pages 130-137, Stroudsburg, PA, USA, 1991. Association for Computational Linguistics. doi : 10.3115/981344.981361. URL `https://doi.org/10.3115/981344.981361`.

Béatrice Daille. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. Thèse en informatique fondamentale, Université de Paris 7, Paris, France, 1994.

Béatrice Daille. Conceptual structuring through term variations. In F. Bond, A. Kohonen, D. Mac Carthy, and A. Villaciencio, editors, *Proceedings of the ACL'2003 Workshop on Multiword Expressions : Analysis, Acquisition, and Treatment*, pages 9-16, 2003.

Béatrice Daille. Building bilingual terminologies from comparable corpora : The ttc termsuite. *The 5th Workshop on Building and Using Comparable Corpora*, page 29, 05 2012.

Ali Darwish. *Terminology and Translation : A Phonological-semantic Approach to Arabic Terminology*. Writescope, Melbourne, Australia, 01 2009. ISBN 978-0-9870709-4-4.

Kareem Darwish, Walid Magdy, and Ahmed Mourad. Language processing for Arabic microblog retrieval. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*, pages 2427-2430. ACM, 2012. ISBN 978-1-4503-1156-4. doi : 10.1145/2396761.2398658. URL http://doi.acm.org/10.1145/2396761.2398658.

S. David and P. Plante. De la nécessité d'une approche morpho-syntaxique dans l'analyse de textes. *Intelligence Artificielle et Sciences Cognitives au Québec*, 3(3) :140-154, 1990.

P. Degoulet and M. Fieschi. *Traitement de l'information médicale : méthodes et applications hospitalières*. Manuels informatiques Masson. Masson, 1991. ISBN 9782225825149. URL https://books.google.fr/books?id=hhbjAAAACAAJ.

Rodolfo Delmonte. Vest - venice symbolic tagger. *Intelligenza Artificiale*, Anno IV(2) :26-27, 2007.

Mona Diab, Kadri Hacioglu, and Daniel Jurafsky. Automatic processing of modern standard arabic text. In Abdelhadi Soudi, Antal van den Bosch, and Günter Neumann, editors, *Arabic Computational Morphology : Knowledge-based and Empirical Methods*, pages 159-179, Dordrecht, 2007. Springer Netherlands. ISBN 978-1-4020-6046-5. doi : 10.1007/978-1-4020-6046-5_9. URL https://doi.org/10.1007/978-1-4020-6046-5_9.

Lee Raymond Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3) :297--302, July 1945. URL http://www.jstor.org/pss/1932409.

Patrick Drouin. *Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés*. PhD thesis, Université de Montréal, 2002.

Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *COMPUTATIONAL LINGUISTICS*, 19(1) :61-74, 1993.

Helge Dyvik. Translations as semantic mirrors : From parallel corpus to wordnet. In *Proceedings of the Workshop Multilinguality in the Lexicon II at the 13th Biennial European Conference on Artificial Intelligence*, ECAI'98, pages 24-44, Brighton, UK, 1998.

Hervé Déjean and Éric Gaussier. Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica*, 19(4) :1-22, 01 2002.

Lois L. Earl. Experiments in automatic extracting and indexing. *Information Storage and Retrieval*, 6(4) :313 - 330, 1970. ISSN 0020-0271. doi : https://doi.org/10.1016/0020-0271(70)90025-2. URL `http://www.sciencedirect.com/science/article/pii/0020027170900252`.

Peter G. Emery. Towards the creation of a unified scientific terminology in arabic. In Barbara Snell, editor, *Term banks for tomorrow's world : Translating and the Computer 4*, London : Aslib, 11-12 November 1982.

Chantal Enguehard and Laurent Pantera. Automatic natural acquisition of a terminology. *Journal of Quantitative Linguistics*, 2(1) :27-32, 1995.

Xiaorong Fan, Nobuyuki Shimizu, and Hiroshi Nakagawa. Automatic extraction of bilingual terms from a Chinese-Japanese parallel corpus. In *Proceedings of the 3rd International Universal Communication Symposium*, IUCS '09, pages 41-45, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-641-0. doi : 10.1145/1667780.1667789. URL `http://doi.acm.org/10.1145/1667780.1667789`.

Mohamed Abdel Fattah, Fuji Ren, and Shingo Kuroiwa. Machine transliteration. In *Proceedings of the 20st Pacific Asia Conference on Language, Information and Computation, PACLIC 20, Huazhong Normal University*, Wuhan, China, November 2006. URL `http://aclweb.org/anthology/Y/Y06/Y06-1050.pdf`.

Fathi Fawi and Rodolfo Delmonte. Italian-Arabic domain terminology extraction from parallel corpora. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*, page 130. Accademia University Press, 2015.

Katerina T. Frantzi and Sophia Ananiadou. Automatic term recognition using contextual cues. In *In Proceedings of 3rd DELOS Workshop*, 1997.

Katerina T. Frantzi, Sophia Ananiadou, and Hideki Mima. Automatic recognition of multi-word terms : the C-Value/NC-Value method. *International Journal on Digital Libraries*, 3 (2) :115-130, 2000.

Pascale Fung. A statistical view on bilingual lexicon extraction : From parallel corpora to non-parallel corpora. In *Machine Translation and the Information Soup*, pages 1-17, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg.

Pascale Fung and Kathleen Mckeown. Finding terminology translations from non-parallel corpora. *Proceedings of the 5th Annual Workshop on Very Large Corpora*, page 192–202., 07 1997.

Pascale Fung and Lo Yuen Yee. An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*, COLING '98, pages 414-420, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics. doi : 10.3115/980451.980916. URL `https://doi.org/10.3115/980451.980916`.

Daniel Gouadec. *Terminologie : constitution des données*. AFNOR gestion. AFNOR, 1990. ISBN 9782124848119. URL `https://books.google.tn/books?id=cRrnAAAAMAAJ`.

Spence Green and Christopher D. Manning. Better Arabic parsing : Baselines, evaluations, and analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 394-402, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL `http://www.aclweb.org/anthology/C10-1045`.

Nizar Habash. *Introduction to Arabic Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2010.

Nizar Habash, Owen Rambow, and Ryan Roth. 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)*, 01 2009.

Nizar Habash, Owen Rambow, and Ryan Roth. *MADA+TOKAN Manual*, June 2010. CCLS-10-01.

Lamia Belguith Hadrich and Nouha Chaaben. Analyse et désambiguïsation morphologiques de textes arabes non voyellés. In *Actes de la 13 ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN'06)*, pages 493-501, Leuven, Belgique, 2006.

Thierry Hamon and Natalia Grabar. Adaptation of cross-lingual transfer methods for the building of medical terminology in Ukrainian. In *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING2016)*, LNCS. Springer, April 2016.

Thierry Hamon, Christopher Engström, and Sergei Silvestrov. Term ranking adaptation to the domain : genetic algorithm based optimisation of the C-Value. In Springer, editor, *Proceedings of PolTAL 2014 - Advances in Natural Language Processing*, volume 8686 of *LNAI*, pages 71-83, September 2014.

Hassan Hamzé. Terminologie grammaticale arabe et terminologie linguistique moderne. *Synergies Tunisie*, 2 :39-54, 2010.

Valérie Hanoka. *Extraction et Complétion de Terminologies Multilingues - Contributions à l'analyse, à l'algèbre et à la combinatoire des endomorphismes sur les espaces de séries*. Theses, Université Paris Diderot (Paris 7), Juillet 2015.

Masamichi Ideue, Kazuhide Yamamoto, Masao Utiyama, and Eiichiro Sumita. A comparison of unsupervised bilingual term extraction methods using phrase tables. In *Proc. MT Summit XIII*, Xiamen, 2011.

ISO 1087 :1990. Terminology - Vocabulary, 1990.

ISO 233-2 (1993). Translittération de l'arabe, 2010.

ISO 704 :2009(fr). Travail terminologique - Principes et méthodes, 2009.

Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1) :11-21, 1972.

John S. Justeson and Slava M. Katz. Technical terminology : some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1) :9-27, 1995.

Kyo Kageura and Bin Umino. Methods of automatic term recognition - a review. *Terminology*, 3(2) :259-289, 1996.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses : Open source toolkit for statistical machine translation. In *ACL*, 2007.

Georgios Kontonatsios, Claudiu Mihaila, Ioannis Korkontzelos, Paul Thompson, and Sophia Ananiadou. A hybrid approach to compiling bilingual dictionaries of medical terms from parallel corpora. In *Statistical Language and Speech Processing*, volume 8791 of *LNCS*, pages 57-69, 2014.

Yuliya Korenchuk. Extraction terminologique : vers la minimisation de ressources. In *Traitement Automatique des Langues Naturelles (TALN)*, Marseille, France, 07 2014. doi : 10.13140/2.1.3125.0565.

Ioannis Korkontzelos, Ioannis P. Klapaftis, and Suresh Manandhar. Reviewing and evaluating automatic term recognition techniques. In Bengt Nordström and Aarne Ranta, editors, *6th International Conference on NLP (GoTAL 2008)*, number 5221 in LNAI, pages 248-259. Springer, August 2008.

Pierre Lafon. Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1 :127-165, 01 1980. doi : 10.3406/mots.1980.1008.

Wiem Lahbib, Ibrahim Bounhas, and Bilel Elayeb. Arabic-English domain terminology extraction from aligned corpora. In Robert Meersman, Hervé Panetto, Tharam Dillon, Michele Missikoff, Lin Liu, Oscar Pastor, Alfredo Cuzzocrea, and Timos Sellis, editors, *On the Move to Meaningful Internet Systems (OTM 2014 Conferences, Confederated International Conferences : CoopIS, and ODBASE 2014, Amantea, Italy, October 27-31, 2014, Proceedings)*, pages 745-759. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014. ISBN 978-3-662-45563-0. doi : 10.1007/978-3-662-45563-0_46. URL `http://dx.doi.org/10.1007/978-3-662-45563-0_46`.

Geoffrey Leech, Roger Garside, and Michael Bryant. Claws4 : The tagging of the british national corpus. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 1*, COLING '94, pages 622--628, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics. doi : 10.3115/991886.991996. URL `https://doi.org/10.3115/991886.991996`.

Abdelkader El Mahdaouy, Saïd El Alaoui Ouatik, and Éric Gaussier. A study of association measures and their combination for Arabic MWT extraction. In *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence (TIA 2013)*, 2013.

Yusney Marrero García, Paloma Moreda Pozo, and Rafael Muñoz-Guillena. Pattern construction for extracting domain terminology. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 420--426. INCOMA Ltd. Shoumen, BULGARIA, 2015. URL `http://aclweb.org/anthology/R15-1055`.

Elizabeth Marshman, Julie L. Gariépy, and Charissa Harms. Helping language professionals relate to terms : Terminological relations and termbases. *Journal of Specialised Translation*, 18, 2012.

Rania Massoud. La terminologie au Liban : réalités et défis. *Annales de l'Institut de langues et de traduction (ILT)*, 10, 2003.

Diana Maynard and Sophia Ananiadou. Identifying terms by their family and friends. In *Proceedings of COLING 2000*, pages 530-536, Saarbrucken, Germany, 2000.

Ryan McDonald, Slav Petrov, and Keith Hall. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 62-72, 2011.

Houari Meyahi. Les difficultés de passage de la terminologie linguistique du français vers l'arabe. *Synergies Algérie*, 20 :93-107, 2013.

Hamdy Mubarak and Ahmed Abdelali. Arabic to English person name transliteration using twitter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, may 2016. ISBN 978-2-9517408-9-1.

Rogelio Nazar, Jorge Vivaldi, and Leo Wanner. Automatic taxonomy extraction for specialized domains using distributional semantics. *Terminology*, 18(2) :188-225, 2012.

Wafa Neifar and Ahmed Ben Ltaief. Acquisition terminologique en arabe : Etat de l'art. In *Actes de JEP-TALN-RECITAL 2016*, volume 3, pages 1-12, Paris, France, Juillet 2016.

Wafa Neifar, Thierry Hamon, Pierre Zweigenbaum, Mariem Ellouze Khemakhem, and Lamia Hadrich Belguith. Adaptation of a term extractor to Arabic specialised texts : First expe-

riments and limits. In *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING2016)*, LNCS. Springer, April 2016a.

Wafa Neifar, Thierry Hamon, Pierre Zweigenbaum, Mariem Ellouze Khemakhem, and Lamia Hadrich Belguith. Impact de l'agglutination dans l'extraction de termes en arabe standard moderne. In *Actes de JEP-TALN-RECITAL 2016*, volume 2, pages 467-474, Paris, France, Juillet 2016b.

Wafa Neifar, Thierry Hamon, Pierre Zweigenbaum, Meriem Ellouze Khemakhem, and Lamia Hadrich-Belguith. Détection des couples de termes translittérés à partir d'un corpus parallèle anglais-arabe. In *Actes de la 25ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2018)*, Rennes, France, Mai 2018. Article court.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1) :19-51, 2003.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. MADAMIRA : A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.

S. Pavel, Canada. Bureau de la traduction. Direction de la terminologie et de la normalisation, D. Nolet, and Canada. Travaux publics et services gouvernementaux Canada. Direction de la terminologie et de la normalisation. *Précis de terminologie*. Terminologie et normalisation, Bureau de la traduction, 2002. ISBN 9780660966670. URL `https://books.google.tn/books?id=nBW0AAAACAAJ`.

Maria Teresa Pazienza, Marco Pennacchiotti, and Fabio Massimo Zanzotto. Terminology extraction : An analysis of linguistic and statistical approaches. In Spiros Sirmakessis, editor, *Knowledge Mining*, volume 185 of *Studies in Fuzziness and Soft Computing*, pages 255-279. Springer Berlin Heidelberg, 2005. URL `http://dx.doi.org/10.1007/3-540-32394-5_20`.

Alain Rey. *La Terminologie : noms et notions*. Que Sais-Je ? Presses Univ. de France, 1979. ISBN 9782130360476. URL `https://books.google.fr/books?id=7jJKAAAAYAAJ`.

Michael Riley, Cyril Allauzen, and Martin Jansche. Openfst : An open-source, weighted finite-state transducer library and its applications to speech and language. In *Proceedings of the North American Chapter of the Association for Computational Linguistics -- Human Language Technologies (NAACL HLT)*, pages 9-10, 01 2009. Tutorials.

Guy Rondeau. *Introduction à la terminologie*. Gaëtan Morin, 1984. ISBN 9782891051378. URL `https://books.google.fr/books?id=fEp8AAAACAAJ`.

Ryan Roth, Owen Rambow, Nizar Habash, Mona Diab, and Cynthia Rudin. Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In *Proceedings of ACL-08 : HLT, Short Papers*, pages 117-120, Columbus, Ohio, June 2008. URL `http://www.aclweb.org/anthology/P/P08/P08-2030`.

Houda Saadane and Nasredine Semmar. Utilisation de la translittération arabe pour l'amélioration de l'alignement de mots à partir de corpus parallèles français-arabe. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*, pages 127-140. ATALA/AFCP, 2012. URL `http://www.aclweb.org/anthology/F12-2010`.

Juan C. Sager. *A Practical Course in Terminology Processing*. J. Benjamins Publishing Company, 1990. ISBN 9789027220769.

Naomi. Sager, Carol Friedman, and Margaret S. Lyman. *Medical Language Processing : Computer Management of Narrative Data*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1987. ISBN 0201168103.

Doaa Samy, Antonio Moreno-Sandoval, Conchi Bueno-Díaz, Marta Garrote-Salazar, and José M. Guirao. Medical term extraction in an Arabic medical corpus. In *Proceedings of LREC'12*, pages 640-645, May 2012.

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In Daniel Jones and Harold Somers, editors, *New Methods in Language Processing Studies in Computational Linguistics*, 1997.

Nasredine Semmar and Houda Saadane. Etude de l'impact de la translittération de noms propres sur la qualité de l'alignement de mots à partir de corpus parallèles français-arabe). In *Traitement Automatique des Langues Naturelles (TALN 2014)*, pages 268-279, Marseille, France, Juillet 2014.

Tarek Sherif and Grzegorz Kondrak. Bootstrapping a stochastic transducer for Arabic-English transliteration extraction. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 864-871, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P07-1109.

Harold Somers. Bilingual parallel corpora and language engineering. In *Proceedings of the Workshop on Language Engineering for South-Asian Languages*, 2001.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 2003*, pages 252-259, 2003.

Jorge Vivaldi and Horacio Rodríguez. Evaluation of terms and term extraction systems : A practical approach. *Terminology*, 13 :225-248, 11 2007.

Ivan Vulic and Marie-Francine Moens. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *ACL*, 2015.

S.E. Wright and G. Budin. *Handbook of Terminology Management*. Number vol. 2 in Application-oriented Terminology Management. J. Benjamins, 2001. ISBN 9789027221551. URL https://books.google.fr/books?id=UYm7XvBXm7QC.

Henrik.R Wulff. The language of medicine. *Journal of the Royal Society of Medicine*, 97 : 187-188, 4 2004.

Eugen Wüster. *Einführung in die Allgemeine Terminologielehre und Terminologische Lexikographie*. Schriftenreihe der Technischen Universität Wien. Springer Vienna, 1979. ISBN 9783211815427.

Behrang Qasemi Zadeh and Siegfried Handschuh. Evaluation of technology term recognition with random indexing. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014.

Ziqi Zhang, José Iria, Christopher Brewster, and Fabio Ciravegna. A comparative evaluation of term recognition algorithms. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*, 2008. URL http://www.lrec-conf.org/proceedings/lrec2008/summaries/538.html.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre :Méthodes d'acquisition terminologique en arabe : Application au domaine médical

Mots clés : acquisition terminologique, transfert multilingue, langue arabe, corpus parallèle, translittération

Résumé : L'objectif de cette thèse est de lever les verrous que constituent le manque de disponibilité de ressources ou d'outils TAL pour la langue arabe dans les domaines de spécialité en proposant des méthodes permettant l'extraction de termes à partir de textes en arabe standard moderne. Dans ce contexte, nous avons d'abord construit un corpus parallèle anglais-arabe dans un domaine de spécialité. Il s'agit d'un ensemble de textes médicaux produits par la bibliothèque nationale de médecine américaine *(NLM)*. Par la suite, nous avons proposé des méthodes d'acquisition terminologique, permettant d'extraire des termes ou d'acquérir des relations entre ces termes, pour la langue arabe en se basant sur : i) adaptation d'un extracteur terminologique existant pour la langue française ou anglaise, ii) l'exploitation de la translittération des termes anglais en caractères arabes et iii) l'application de la la notion de transfert translingue. Appliqué au niveau terminologique, le transfert consiste à mettre en œuvre un processus d'extraction de termes ou d'acquisition de relations entre termes sur des textes d'une langue source (ici, le français ou l'anglais) puis à transférer les informations extraites sur des textes d'une langue cible (ici, l'arabe standard moderne) pour ainsi identifier le même type d'informations terminologiques. Nous avons évalué les listes de termes monolingues et bilingues obtenues lors des différentes expériences que nous avons réalisées, suivant une méthode transparente, directe et semi-automatique : les termes candidats extraits sont confrontés à une terminologie de référence avant d'être vérifiés manuellement. Cette évaluation suit un protocole que nous avons proposé.

Title : Terminology acquisition methods in Arabic: Application to the medical domain

Keywords : Terminology Acquisition, Multilingual Transfer, Arabic Language, Parallel Corpora, Transliteration

Abstract : The goal of this thesis is to reduce the lack of available resources and NLP tools for Arabic language in specialised domains by proposing methods allowing the extraction of terms from texts in Modern Standard Arabic. In this context, we first constructed an English-Arabic parallel corous in a specific domain. It is a set of medical texts produced by the US National Library of Medicine *(NLM)*. Thereafter, we have proposed terminological acquisition methods, to extract terms or acquire relations between these terms, for Arabic based on: i) the adaptation of an existing terminology extractor for French or English, ii) the transliteration of English terms in Arabic characters and iii) cross-lingual transfer. Applied at the terminological level, transfer aims to implement a process of term extraction or relationship acquisition between terms in the texts of a source language (here, French or English) and then to transfer the extracted information to target language texts (in this case, Modern Standard Arabic), thereby identifying the same type of terminological information. We have evaluated the monolingual and bilingual term lists that we have obtained by the experiments we carried out, according to a transparent, direct and semi-automatic method: the extracted term candidates are confronted with a reference terminology before being validated manually. This evaluation follows a protocol that we proposed.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France