# Acquisition terminologique en arabe : État de l'art

Wafa Neifar[1,2] Ahmed Ben Ltaief[2]
(1) LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
(2) Laboratoire MIRACL, Université de Sfax, Tunisie
neifar@limsi.fr , ahmedbenltaief92@gmail.com

RÉSUMÉ

L'acquisition terminologique est une tâche indispensable pour l'accès aux informations présentes dans les corpus de spécialité. Il s'agit d'une part, d'identifier et d'extraire des termes, et d'autre part, de structurer ces termes à l'aide de méthodes d'acquisition de relations sémantiques. Dans cet article, nous nous intéressons à l'acquisition terminologique sur des textes en arabe standard moderne (MSA). Nous réalisons tout d'abord, un état de l'art décrivant les méthodes d'extraction de termes sur cette langue ainsi que les approches proposées pour la reconnaissance de relations sémantiques entre termes issus. Après avoir présenté quelques corpus de spécialité et ressources terminologiques disponibles en MSA que nous avons identifiés, nous décrivons nos premières pistes de travail.

ABSTRACT

**Terminological acquisition on MSA : State of the art**

Terminological acquisition is required to access the information in the specialised texts. It first concerns the identification and the extraction terms, and then, the structuration of these terms with methods dedicated to the acquisition of semantic relations. In this paper, we focus on the terminological acquisition on Modern Standard Arabic (MSA) texts. We start by a state of the art of the methods of term extraction and approaches for recognizing semantic relations between terms on this language. Then, after a description of some MSA specialised corpora and terminological resources availables, we prersention our first proposition for the term extraction and the relation acquisition.

MOTS-CLÉS : Terminologie, corpus de spécialité, Arabe standard moderne, Extraction de termes, Acquisition de relations sémantiques.

KEYWORDS : Terminology, Specialized corpora, Modern Standard Arabic, Terms extraction, Semantic relation acquisition.

## 1 Introduction

L'arabe standard moderne (MSA) est une langue sémitique dont l'origine est l'arabe classique c'est-à-dire la langue du Coran. Il correspond à une évolution moderne de l'arabe classique. Ainsi, même s'il existent de nombreux dialectes, les locuteurs dont la langue arabe est la langue maternelle peuvent communiquer entre eux. L'arabe 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 produits dans cette langue. Il est donc important de disposer de systèmes de gestion terminologique. L'aménagement terminologique est nécessaire dans de nombreux autres

domaines comme l'agriculture, la géologie, la protection de l'environnement ou le droit, bien que celui-ci peut varier d'un pays à l'autre (Massoud, 2003). Elle n'est cependant pas utilisée dans tous les domaines de spécialité. Ainsi en médecine, la langue utilisée lors de la pratique et de l'enseignement est la langue Française ou anglaise (Samy *et al.*, 2012). Or, la compréhension des notions spécialisées par le grand public, et notamment les patients, est primordiale.

Dans ce contexte, les méthodes d'acquisition terminologique jouent un rôle important. L'objectif à terme de notre travail est de proposer des méthodes permettant la constitution de ressources terminologiques en MSA. Aussi, dans cet article, nous présentons un état de l'art des travaux d'acquisition terminologique en arabe. Nous passons en revue les approches d'extraction terminologique (section 2.1) et de relations sémantiques entre ces termes sur cette langue (section 2.2). La section 3 est consacrée à la présentation de quelques corpus de spécialité disponibles ainsi que les ressources terminologiques pouvant être utiles dans notre travail. Nous proposons ensuite quelques résultats sur l'adaptation de l'extracteur de termes YATEAau MSA (section 4.1) mais aussi, une première analyse des phénomènes nous permettant d'identifier des relations sémantiques entre termes arabes (section 4.2).

# 2 Acquisition terminologique en MSA

Depuis les années 90, de nombreuses approches ont été proposées pour aider à la constitution de ressources terminologiques à partir de textes de spécialité (articles scientifiques, documentations techniques, textes juridiques, etc.) (Cabré *et al.*, 2001 ; Pazienza *et al.*, 2005). Il s'agit d'une part d'identifier et d'extraire des termes à partir de textes (Pazienza *et al.*, 2005 ; Q. Zadeh & Handschuh, 2014) et, d'autre part de mettre en relation ou de regrouper ces termes extraits (Marshman *et al.*, 2012). Bien que les méthodes distributionnelles ou de classification supervisées peuvent être utilisées (Nazar *et al.*, 2012 ; Zadeh & Handschuh, 2014 ; Conrado *et al.*, 2013), les approches principalement utilisées s'appuient à la fois sur une description linguistique du processus d'extraction et des filtres statistiques (Bourigault, 1993 ; Daille, 2003 ; Drouin, 2002 ; Aubin & Hamon, 2006).

Avec l'intérêt croissant pour le traitement automatique de la langue arabe, des méthodes d'acquisition terminologiques à partir de textes en MSA ont également été proposées. Néanmoins, la plupart des travaux utilise des approches similaires à ceux réalisées sur l'anglais ou le français.

## 2.1 Extraction de termes

La complexité de la langue arabe est une difficulté importante lors de la conception de méthodes d'extraction de termes. En effet, elle demande aux approches traditionnelles d'acquisition terminologique de prendre 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). Á partir de ces observations, Boulaknadel *et al.* (2008) considèrent que toutes ces particularités empêchent la mise en œuvre d'approches statistiques et que l'extraction terminologique doit surtout s'appuyer sur des méthodes linguistiques. Dans ce contexte, l'approche proposée par les auteurs s'appuie principalement sur une description précise de la formation et de la variation des termes de manière similaire à des travaux précédents sur le français et l'anglais (Daille, 2003).

En général, les travaux existant proposent de combiner des règles s'appuyant sur des critères linguistiques pour décrire les termes à extraire, avec des informations statistiques pour sélectionner les

termes candidats les plus probables. Ainsi, Bounhas & 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 la mesure d'association LLR (*Log-Likelihood Ratio*) est appliqué sur les résultats ambigus de la première étape pour sélectionner la meilleure solution. AlKhatib & 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 (Maynard & Ananiadou, 2000) 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 d'environnement, collectés à partir des sites Web. De même, 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 extraire automatiquement les termes simples et complexes du domaine. Un corpus électronique contenant l'Arabe Classique (CA) et l'Arabe Standard Moderne (MSA), collecté à partir des archives de journaux islamiques et des sites islamiques, est utilisé pour l'évaluation de l'approche. Dans ce travail, le TF*IDF est utilisé pour trier les termes simples en fonction de leur *termhood*, et plusieurs mesures (information mutuelle, Kappa, $\chi^2$, T-test, Piatersky-Shapiro et l'agrégation des rang) sont utilisées pour calculer le degré d'association de leur composants (*unithood*).

A l'exception de (Bounhas *et al.*, 2011), les méthodes précédentes utilisent des approches très similaires à celles proposées pour l'anglais ou le français sans tenir compte des particularités linguistiques de l'arabe comme les ambiguïtés morphologiques et syntaxiques, la non-voyellation ou l'agglutination. De plus, 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, 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 collections de textes en arabe issus de domaine de spécialité.

Outre les problèmes liés à l'évaluation, les systèmes issus des travaux présentés ci-dessus ne sont pas librement accessibles. Afin de palier cet inconvénient, nous souhaitons développer notre approche pour l'extraction de termes en adaptant un système existant. Pour cela, nous avons choisi de nous appuyer sur l'extracteur de termes YaTeA[1] (Aubin & Hamon, 2006) en définissant les règles d'extraction spécifiques au MSA.

## 2.2 Acquisition de relations sémantiques

Á notre connaissance, assez peu de travaux se sont intéressés à l'acquisition de relations sémantiques entre termes à partir de textes spécialisés en MSA. Aussi, nous présentons également ici des travaux portant sur l'identification de relations entre entités nommées. En effet, ceux-ci pourraient être mise en œuvre pour acquérir des relations entre termes.

Plusieurs travaux exploitent la morphologie de la langue arabe pour acquérir des relations entre

[1] librement disponible à l'adresse suivante `http://search.cpan.org/~thhamon/Lingua-YaTeA/`

termes. Boulaknadel *et al.* (2008) s'appuient sur une description des variations morpho-syntaxiques et syntaxiques pour identifier des relations entre termes comme بئر نفطي (puits pétrolier) / من النفط بئر (puits de pétrole) ou التكوين للتربة (composition du sol) / التكوين المستمر للتربة (composition permanente du sol). Avec un objectif similaire, Belkredim & Sebai (2009) proposent une formalisation des phénomènes de dérivation dans une ontologie. L'objectif de ce travail est de définir des règles de dérivation qui devraient permettre d'identifier des verbes dérivés de noms.

L'acquisition des relations de sémantiques peut également s'appuyer sur la structure syntaxique des termes. Ainsi, Lahbib *et al.* (2013) utilisent les dépendances syntaxiques au sein de groupes nominaux pour identifier des relations de causalité ou d'association. Un calcul de similarité sémantique basé notamment sur le LLR entre les termes mis en relation permet de retenir les plus probables. L'évaluation de la méthode proposée est réalisée sur des textes religieux voyellés (Hadith) afin d'éviter les problèmes d'ambiguïté généralement rencontrés sur des textes non voyellés. Les résultats obtenus sont ainsi quasi-parfaits (entre 97% et 100% de précision) mais il est difficile d'évaluer la portabilité de l'approche étant donné que la plupart des textes en MSA sont non voyellés. De même, (Ben Hamadou *et al.*, 2010) décrivent les relations sémantiques que peuvent entretenir des entités nommées entre elles à l'aide de transducteurs dans la plate-forme linguistique NooJ. L'application de l'approche proposée sur des pages Wikipedia permet d'obtenir une F-mesure de 70%.

Enfin, les méthodes d'apprentissage peuvent également identifier des règles d'association caractéristiques entre relations entre paires d'entités nommées (Boujelben *et al.*, 2013). Celles-ci sont combinées à un filtrage basé sur la description des associations produites, afin de réduire le grand nombre de règles produites. L'évaluation sur un corpus de phrases issues de différents articles journalistiques arabes permet d'obtenir une F-mesure de 60%. Afin d'améliorer la méthode précédente, des algorithmes génétiques sont utilisés lors de l'étape de filtrage. L'évaluation est cette fois réalisée sur le corpus ANERCorp (Benajiba *et al.*, 2007) et permet d'obtenir une F-mesure de 66%.

Notons également l'approche proposée par Faruqui & Kumar (2015), bien qu'elle ne soit pas actuellement mise en œuvre sur l'arabe. Celle-ci exploite la notion de transfert pour acquérir des relations :

- dans un premier temps, chaque phrase est traduite dans une langue cible, ici, l'anglais grâce à l'API Google Translate. L'utilisation de ce système de traduction automatique permet d'éviter l'inconvénient de disposer de corpus parallèle nécessaire à ce type de méthode et fournit un alignement des phrases au niveau du mot.

- Les relations sont ensuite identifiées sur le corpus anglais en utilisant le système OLLIE (Mausam *et al.*, 2012).

- les relations identifiées sont projetées sur la langue source à l'aide de l'alignement pour acquérir des relations dans la langue source.

L'évaluation des performances de cette approche est réalisée sur trois langues : le français, l'hindi et le russe. L'acquisition des relations à partir de phrases extraites de Wikipédia montre des résultats variables : 81.6%, 64.9% et 63.5%.

# 3 Matériel disponible en domaine de spécialité

## 3.1 Corpus de spécialité en MSA

La mise en œuvre et l'évaluation des méthodes d'extraction de termes ou d'acquisition de relations sémantiques nécessite de disposer un corpus arabe de spécialité. Cependant, assez peu de corpus dans des domaines spécifiques sont librement disponibles pour la langue arabe. En effet, alors que plusieurs travaux mentionnent l'utilisation d'un corpus en arabe de pages Web issues du domaine d'environnement (Bounhas & Slimani, 2009 ; AlKhatib & Badarneh, 2010), du domaine médical, (Al-Sulaiti & Atwell, 2006) ou en lien avec la religion Abed *et al.* (2013), ces corpus ne sont pas actuellement accessibles.

Aussi, contrairement à la plupart des travaux d'extraction de termes en arabe, nous ne souhaitions pas travailler sur des données textuelles issues de sites Web ou de forums, car la qualité terminologique est difficilement vérifiable. Aussi, nous avons constitué notre corpus à partir de textes produits par la *National Library of Medicine* (NLM) et disponibles en ligne sur MedlinePlus[2]. Ces documents sont des brochures de quelques pages à destination des patients. Ces brochures apportent des informations sur des problèmes médicaux, dérivent les conditions de réalisation d'examen, fournissent des conseils de comportement face à une maladie, ou pour l'amélioration du bien-être. Si l'anglais est la langue source, ces documents sont traduits dans de nombreuses langues cibles comme le français ou l'arabe. Actuellement, nous disposons de 504 textes parallèles arabe/français/anglais au format PDF. Contrairement aux documents en anglais et en français, la conversion au format texte des documents en arabe, en vue de réaliser des traitements automatiques posent un certain nombre de problèmes rendant la tâche de nettoyage et de normalisation très coûteuse en temps (section 3.1.1). Aussi, pour les expériences dans cet article, nous avons dû nous limiter à un corpus de 30 textes médicaux en arabe, composé 15 532 mots.

### 3.1.1 Pré-traitements des textes arabes

Lors de la conversion au format texte et au nettoyage des documents MedlinePlus en langue arabe, nous avons constaté plusieurs problèmes de codage de caractères dont certains sont déjà décrits dans (Habash, 2010). Ces problèmes entraînent des difficultés de mise en œuvre des approches ou des outils du TAL, par exemple, lors de la projection de dictionnaires. Nous avons donc réalisé des pré-traitements spécifiques consistant à nettoyer et normaliser les textes du corpus.

Un premier problème concerne les caractères spéciaux UTF-8 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. Il est donc important de les supprimer afin de pas perturber les traitements basés sur des expressions régulières.

Si ce premier problème est assez simple à corriger, les erreurs graphiques ou orthographiques sont des problèmes majeurs qui sont plus complexes à éliminer et peuvent empêcher la projection correcte de dictionnaire. Il peut s'agir

- du remplacement d'un caractère arabe par un autre arabe. Ainsi le caractère ī est utilisé dans le mot لـ ī ـشا, alors que la forme correcte du mot est مشاكل -- *problèmes*).

[2] `http://www.nlm.nih.gov/medlineplus/languages/all_healthtopics.html`

- du remplacement d'un caractère arabe par un caractère persan qui lui ressemble graphiquement. Par exemple, il est possible de rencontrer le caractère farsi ی (U+06CC, *Arabic Letter Yeh Farsi*) à la place du ى (U+0649, *Alef Maksura*)[3]

  La forme utilisant le caractère initial farsi يحوم et la forme utilisant caractère initial arabe يحوم (*se base*) sont similaires graphiquement lorsque les caractères sont détachés, mais lorsqu'il est dans sa forme attachée, le caractère arabe apparaît différent. On retrouve alors la première forme, erronée, dans des textes arabes.

- de l'inversement des positions de caractères. Il est ainsi possible de rencontrer par exemple la forme (ألوعية au lieu de الأوعية (les vaisseaux) ou الخاليا au lieu de الخلايا) (les cellules).

- de l'inexactitude des formes des caractères. On observe sous la (اسىتنشاقها au lieu de استنشاقها (*l'inhaler*), ou امرتصاصها au lieu امتصاصها) (*son absorption*).

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.

### 3.1.2 Analyse morphologique et étiquetage morpho-syntaxique

De manière générale, l'état de l'art montre qu'il est préférable de disposer de textes étiquetés morpho-syntaxiquement et lemmatisés pour réaliser l'extraction des termes candidats. Parmi l'ensemble des systèmes disponibles, deux outils sont plus largement utilisés : Stanford POS Tagger (Toutanova *et al.*, 2003) et MADA+TOKAN (Habash *et al.*, 2010). Les précédentes comparaisons entre ces deux outils (Green & Manning, 2010 ; Albogamy & Ramsay, 2015) montrent que les résultats obtenus avec MADA+TOKAN sont de meilleure qualité. Nous avons opté pour l'outil MADA+TOKAN qui, outre la qualité de son analyse morphologique, offre une lemmatisation des mots du corpus. MADA+TOKAN fournit également, plusieurs traits morphologiques associés aux catégories grammaticales (nom, verbe, adverbe...) des mots : le genre (masculin, féminin, non applicable, non défini), le nombre (singulier, duel, pluriel, non applicable, non défini), le cas (nominatif, accusatif, génitif, non applicable, non défini) et l'état (indéfini, défini, possesseur/*idafa*, non applicable, non défini).

## 3.2 Ressources terminologiques en MSA

Á notre connaissance, les ressources terminologiques disponibles pour le MSA au format électronique, sont uniquement issues du domaine médical. Il s'agit d'un part de la Classification Internationale des Maladies (CIM -- `http://www.who.int/classifications/icd/en/`), d'autre part, le *Medical Subject Headings* (MeSH -- `http://www.emro.who.int/fr/information-resources/arabic-mesh/`). Ces ressources nécessitent toutefois une préparation afin d'être exploitable dans un contexte applicatif ou pour être utilisées comme une donnée de référence dans le cadre d'une évaluation d'outil d'acquisition terminologique.

[3] La différence graphique est liée à la police de caractères utilisée.

# 4 Méthode d'acquisition de ressources terminologiques pour le MSA

Dans cette section, nous présentons d'une part une première adaptation de l'extracteur de termes YATEAau MSA, et d'autre part, des premières propositions de patrons d'acquisition de relations sémantiques issus d'observation en corpus. Les exemples qui illustrent les approches proposées sont tirés du corpus MedlinePlus.

## 4.1 Extraction de termes

Lors de l'adaptation de YATEAaux 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. Nous faisons également l'hypothèse qu'une analyse morphologique et une étiquette morpho-syntaxique sont associées à chaque mot du texte à traiter.

YATEAeffectuant une analyse syntaxique superficielle du corpus, la première étape consiste à définir les frontières syntaxiques permettant d'identifier les groupes nominaux maximaux. Pour cela, de manière similaire aux autres langues, les pronoms, les ponctuations, 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 (إن, تكون, كان) (il était, elle est, certes), les adverbes (ربما, هنا, فقط) (peut-être, ici, seulement), ou les expressions lexicales ou les motifs (في بعض الأوقات) qui ne doivent pas non plus faire partie des termes.

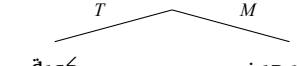
Nous avons également déterminé les étiquettes morpho-syntaxiques qui ne doivent pas figurer au début ou à la fin d'un terme. Il s'agit surtout de prépositions (من (de), إلى (à), بَيْنَ (entre), عند) (lorsque) auxquelles l'analyseur MADA+TOKAN attribue par erreur, la catégorie morpho-syntaxique de nom.

تكون كمية الأكسجين التي يحملها الدم أقل

(A) Texte initial

كمية الأكسجين (quantité d'oxygène) — الدم (le sang)

(B) Étape 1 : Segmentation à l'aide de frontières syntaxiques




كمية (quantité) — الأكسجين (oxygène) — الدم

(C) Étape 2 : Analyse syntaxique des termes

FIGURE 1 – Étape d'extraction des termes sur un extrait en arabe (littéralement : *la quantité d'oxygène dans le sang est moindre*).

Par exemple, dans la phrase présentée à la figure 1(a), les frontières syntaxiques تكون (pseudo-verbe), التي (que/laquelle), يحملها (la transporte) et أقل (moins) permettent d'identifier les syntagmes

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

Après cette étape, des patrons spécifiques sont définis pour l'analyse syntaxique en tête/modifieur des syntagmes nominaux maximaux identifiés précédemment. Il s'agit de séquences d'étiquettes morpho-syntaxiques et de prépositions appliquées récursivement sur les groupes nominaux maximaux pour identifier le rôle syntaxique de leur composants, mais aussi de filtrer les séquences de mots inutiles qui ne pouvent pas être analysées à l'aide des patrons définis. Ces patrons prennent en compte les caractéristiques morphologiques telles le genre et le nombre, mais aussi le cas des constituants comme *al-'iḍāfah* qui marque l'état construit et le génitif et permet d'identifier la tête et le modifieur d'un terme. Par exemple, le patron `noun-m-s-g-d(Modifieur) noun-f-s-n-c(Tête)`[4] permet d'analyser le syntagme maximal كمية الأكسجين (quantité d'oxygène) tel que présenté à la figure 1(c)

## 4.2 Acquisition de relations sémantiques

Pour identifier les relations sémantiques entre les termes extraits à partir des termes extraits, nous avons choisi de proposer des patrons lexico-syntaxiques à l'instar des travaux existants dans d'autres langues (Hearst, 1992 ; Morin, 1999). Nous avons également utilisé la décomposition entre tête/modifieur fournie par YATEA. Nous nous sommes intéressés aux relations de synonymie, d'hyperonymie et de méronymie pour lesquelles nous avons cherché des ancres lexicales, syntaxiques ou typographiques caractéristiques.

**Synonymie** Pour l'acquisition de la relation de synonymie, nous avons construit des patrons en se basant sur des ancres lexicales. Il s'agit soit de verbes, soit d'expressions lexicales. Par exemple, dans la phrase كلمة جلوكوز هي المرادفة لكلمة سكر (*Le mot glucose est le synonyme du mot sucre*), l'ancre lexicale هي المرادفة ل *est le synonyme du* permet d'identifier la relation de synonymie entre les termes جلوكوز (*glucose*) et سكر (*sucre*) La ponctuation peut aussi jouer un role important pour la détection des relations. De ce fait, la présence des parenthèses nous sert aussi, dans certains cas pour relever la synonymie. Certains termes sont suivis par un autre terme marqué entre parenthèses pour présenter le meme concepte autrement. Á partir d'une étude faite sur 19 textes arabes du corpus MedlinePlus, nous avons défini 24 patrons sémantiques pour l'extraction de la synonymie.

**Hyperonymie** La relation d'hyperonyme consiste à spécifier une relation entre un terme général et un terme plus spécifique. Pour l'acquisition de ce type de relation, nous nous sommes appuyés sur l'analyse syntaxique des termes proposée par YATEAadapté au MSA. Nous avons utilisé l'inclusion lexicale (Grabar & Zweigenbaum, 2004) qui permet d'identifier la tête d'un terme comme l'hyperonyme du terme. Ainsi, استئصال الزائدة الدودية (*appendicectomie*) est identifié comme l'hyperonyme de استئصال الزائدة الدودية بالمنظار (*appendicectomie laparoscopique*).

Des patrons peuvent également exploiter des ancres lexicales pour acquérir des relations d'hyperonymie. Par exemple, dans la phrase ضع مرهمًا مضادا حيويًا , مثل النيوسبورين , على الجرح , عند توقف النزيف (*Lorsque le saignement est arreté, mettre la pommade anti-vitale, comme Niuspourin, sur la plaie.*, le mot مثل (*comme*) indique la présence d'une relation d'hyperonymie entre le terme hyperonyme مرهمًا مضادا حيويًا (*une pommade anti-vitale*) et le terme hyponyme النيوسبورين (*Niuspourin*).

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

**Méronymie** La méronymie consiste à préciser une partie de la totalité. Certaines expressions lexicales peuvent être utilisées pour identifier cette relation. Ainsi, dans l'extrait من مفاصل الأصابع كل مفصل (*toutes articulations des articulations des doigts*), la présence de l'expression كل ... من ... (*toutes ... de ...*) permet d'identifier que مفصل (*articulation*) exprime une partie du terme الأصابع مفاصل (*articulations des doigts*).

La présence de la relation de possession (*al-'iḍāfah*), dans un terme est aussi utile pour identifier ce type de relation. Nous exploitons alors l'analyse morpho-syntaxique des termes extraits et notamment l'état construit. Ainsi, à partir de l'exemple précédent, l'état construit du terme الأصابع (*doigts*) peut être utilisé pour acquérir la relation de méronymie avec le terme مفصل (*articulations*)

# 5 Conclusion et perspectives

Dans cet article, nous nous sommes intéressés à présenter un état de l'art sur l'acquisition terminologique pour la langue arabe. Même si plusieurs travaux existent sur cette tâche, les méthodologies proposées ne sont pas reproductibles car elles sont partiellement décrites et les systèmes ne sont pas disponibles. De même, il est difficile d'apprécier les résultats et de les comparer car les méthodes d'évaluation ne sont pas clairement présentées.

Dans ce contexte, nous avons effectué une première adaptation de l'extracteur de termes YATEA afin de pouvoir traiter des textes de spécialité en arabe standard moderne. Il s'agissait de définir le processus d'extraction de termes candidats en s'appuyant sur une description linguistique des termes arabes, mais aussi en prenant en compte les particularités morphologiques de cette langue. Aussi, nous nous sommes intéressés à l'extraction de relations sémantiques entre les termes. Pour cela, nous avons défini des patrons syntaxiques pour chaque type de relation en se basant sur une étude du corpus.

Plusieurs perspectives de travail s'offrent à nous : d'une part, le traitement de la voyellation et de l'agglutination, d'autre part, les marques morphologiques de cas ou le *masdar* pourraient aussi être utilisés pour corriger l'étiquetage morpho-syntaxique dans certains cas bien précis. La prise en compte de ces phénomènes linguistiques permettra d'améliorer l'analyse syntaxique des termes candidats. Aussi, nous envisageons d'évaluer notre travail sur plusieurs corpus issus de différents domaines de spécialité, en fonction de leur disponibilité.

# Références

Abed A. M., Tiun S. & Albared M. (2013). Arabic term extraction using combined approach on islamic document. *Journal of Theoretical & Applied Information Technology*, **58**(3).

Al-Sulaiti L. & Atwell E. (2006). The design of a corpus of Contemporary Arabic. *International Journal of Corpus Linguistics*, **11**(1), 1--36.

Albogamy F. & Ramsay A. (2015). Pos tagging for arabic tweets. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, p. 1--8.

AlKhatib K. & Badarneh A. (2010). Automatic extraction of arabic multi-word terms. In *IMCSIT*, p. 411--418.

Aubin S. & Hamon T. (2006). Improving term extraction with terminological resources. In T. Salakoski, F. Ginter, S. Pyysalo & T. Pahikkala, Eds., *5th International Conference on NLP (FinTAL 2006)*, number 4139 in LNAI, p. 380--387 : Springer.

Belkredim F. Z. & Sebai A. E. (2009). An ontology based formalism for the arabic language using verbs and their derivatives. *Communications of the IBIMA*, **11**, 44--52.

Ben Hamadou A., Piton O. & Fehri H. (2010). Multilingual Extraction of functional relations between Arabic Named Entities using NooJ platform. In Z. Gavriilidou, E. Chadjipapa, L. Papadopoulou & M. Silberztein, Eds., *Nooj 2010 International Conference and Workshop*, p. 192--202, Komotini, Greece : Le Département de Philologie Grecque de l'Université Democritus de Thrace, le Laboratoire de Sémio-Linguistique et Didactique (LASELDI) de l'Université de Franche-Comté et la Maison des Sciences de l'Homme et de l'Environnement Ledoux. 10 pages.

Benajiba Y., Rosso P. & Benedí J. (2007). Anersys : An arabic named entity recognition system based on maximum entropy. In *Proceedings of the 8th Int. Conf. on Comput. Linguistics and Intelligent Text Processing (CICLing-2007)*, number 4394 in LNCS, p. 143--153 : Springer-Verlag.

Boujelben I., Jamoussi S. & Ben A. H. (2013). Enhancing machine learning results for semantic relation extraction. In *Proceedings of the 18th International Conference on Applications of Natural Language to Information Systems (NLDB 2013)*, p. 337--342, Salford, UK : Springer, Berlin Heidelberg.

Boulaknadel S., Daille B. & Aboutajdine D. (2008). A multi-word term extraction program for arabic language. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias, Eds., *Proceedings of the LREC'08*.

Bounhas I., Elayeb B., Evrard F. & Slimani Y. (2011). Organizing contextual knowledge for arabic text disambiguation and terminology extraction. *Knowledge Organization Journal*, **38**(6), 473--490.

Bounhas I., Lahbib W. & Elayeb B. (2014). Arabic domain terminology extraction : A literature review - (short paper). In *OTM 2014 Conferences - Confederated International Conferences : CoopIS, and ODBASE 2014*, p. 792--799, Amantea, Italy.

Bounhas I. & Slimani Y. (2009). A hybrid approach for arabic multi-word term extraction. In *Natural Language Processing and Knowledge Engineering, NLP-KE 2009. International Conference on*, p. 1–8 : IEEE IEEE.

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

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

Conrado M., Pardo T. & Rezende S. (2013). A machine learning approach to automatic term extraction using a rich feature set. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*, p. 16--23, Atlanta, Georgia.

Daille B. (2003). Conceptual structuring through term variations. In F. Bond, A. Kohonen, D. M. Carthy & A. Villaciencio, Eds., *Proceedings of the ACL'2003 Workshop on Multiword Expressions : Analysis, Acquisition, and Treatment*, p. 9--16.

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

Faruqui M. & Kumar S. (2015). Multilingual open relation extraction using cross-lingual projection. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1351--1356, Denver, Colorado : Association for Computational Linguistics.

Grabar N. & Zweigenbaum P. (2004). Lexically-based terminology structuring. *Terminology*, **10**(1), 23--54.

Green S. & Manning C. D. (2010). Better arabic parsing : Baselines, evaluations, and analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, p. 394--402, Beijing, China : Coling 2010 Organizing Committee.

Habash N. (2010). *Introduction to Arabic Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Habash N., Rambow O. & Roth R. (2010). *MADA+TOKAN Manual*. CCLS-10-01.

Hamon T., Engström C. & Silvestrov S. (2014). Term ranking adaptation to the domain : genetic algorithm based optimisation of the C-Value. In Springer, Ed., *Proceedings of PolTAL 2014 -- Advances in Natural Language Processing*, volume 8686 of *LNAI*, p. 71--83.

Hearst M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of 14th International Conference on Computational Linguistics (COLING'92)*, p. 539--545, Nantes, France.

Korkontzelos I., Klapaftis I. P. & Manandhar S. (2008). Reviewing and evaluating automatic term recognition techniques. In B. Nordström & A. Ranta, Eds., *6th International Conference on NLP (GoTAL 2008)*, number 5221 in LNAI, p. 248--259 : Springer.

Lahbib W., Bounhas I., Elayeb B., Evrard F. & Slimani Y. (2013). A hybrid approach for arabic semantic relation extraction. In *26th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2013)*, p. pp. 315--320, St. Pete Beach, Florida's west coast, US : AAAI Press.

Marshman E., Gariépy J. L. & Harms C. (2012). Helping language professionals relate to terms : Terminological relations and termbases. *Journal of Specialised Translation*, **18**.

Massoud R. (2003). La terminologie au liban : réalités et défis. *Annales de l'Institut de langues et de traduction (ILT)*, **10**.

Mausam, Schmitz M., Soderland S., Bart R. & Etzioni O. (2012). Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 523--534, Jeju Island, Korea : Association for Computational Linguistics.

Maynard D. & Ananiadou S. (2000). Identifying terms by their family and friends. In *Proceedings of COLING 2000*, p. 530--536, Saarbrucken, Germany.

Morin E. (1999). Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique. *Traitement Automatique des Langues*, **40**(1).

Nazar R., Vivaldi J. & Wanner L. (2012). Automatic taxonomy extraction for specialized domains using distributional semantics. *Terminology*, **18**(2), 188--225.

Pazienza M. T., Pennacchiotti M. & Zanzotto F. (2005). Terminology extraction : An analysis of linguistic and statistical approaches. In S. Sirmakessis, Ed., *Knowledge Mining*, volume 185 of *Studies in Fuzziness and Soft Computing*, p. 255--279. Springer Berlin Heidelberg.

Q. Zadeh B. & Handschuh S. (2014). The acl rd-tec : A dataset for benchmarking terminology extraction and classification in computational linguistics. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*, p. 52--63, Dublin, Ireland.

Samy D., Moreno-Sandoval A., Bueno-Díaz C., Garrote-Salazar M. & Guirao J. M. (2012). Medical term extraction in an arabic medical corpus. In *Proceedings of LREC'12*.

Toutanova K., Klein D., Manning C. & Singer Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 2003*, p. 252--259.

Zadeh B. Q. & Handschuh S. (2014). Evaluation of technology term recognition with random indexing. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.