RENAM: Système de Reconnaissance des Entités Nommées Amazighes

Meryem Talha[1] Siham Boulaknadel[1, 2] Driss Aboutajdine[1]
(1) LRIT, Unité Associée au CNRST (URAC 29), Faculté des Sciences, Mohammed V-Agdal, Rabat, Maroc
(2) IRCAM, Avenue Allal El Fassi, Madinat Al Irfane, Rabat-Instituts, Maroc
meriem.talha@gmail.com, boulaknadel@ircam.ma, aboutaj@fsr.ac.ma

Résumé. La reconnaissance des Entités Nommées (REN) en langue amazighe est un prétraitement potentiellement utile pour de nombreuses applications du traitement de la langue amazighe. Cette tâche représente toutefois un sévère challenge, compte tenu des particularités de cette langue. Dans cet article, nous présentons le premier système d'extraction d'entités nommées amazighes (RENAM) fondé sur une approche symbolique qui utilise le principe de transducteur à états finis disponible sous la plateforme GATE.

Abstract. Named Entity Recognition (NER) for Amazigh language is a potentially useful pretreatment for many processing applications for the Amazigh language. However, this task represents a tough challenge, given the specificities of this language. In this paper, we present (NERAM) the first named entity system for the Amazigh language based on a symbolic approach that uses linguistic rules built manually by using an information extraction tool available within the platform GATE.

Mots-clés : Reconnaissance des entités nommées (REN), Langue Amazighe, Règles d'annotation, JAPE, GATE.
Keywords: Named Entities Recognition (NER), Amazigh Language, Annotation Rules, JAPE, GATE.

1 Introduction

La langue amazighe fait partie des langues chamito-sémitiques ou encore appelé afro-asiatiques (Cohen 2007, Chaker 1989) qu'on soit au Maroc ou ailleurs. La langue Amazighe est maintenant une langue qui possède tous ses attributs: dotée d'une graphie officielle, un codage propre dans le standard Unicode, une grammaire, une orthographe et un vocabulaire très riche.

La tâche REN est une sous tâche du domaine d'extraction d'information consistant à identifier et à catégoriser certaines expressions linguistiques autonomes et mono-référentielles (Ehrmann, 2008). Certaines langues ont éveillé beaucoup d'intérêt, notamment à travers les campagnes d'évaluation telles que CONLL (Tjong Kim Sang, 2002) pour l'espagnol et l'allemand, MUC (Grishman et Sundheim, 1996) pour l'anglais et le japonais, et ESTER (Galliano et al., 2009) pour le français. Toutefois, l'Amazighe étant une langue peu dotée en terme de ressources linguistiques informatisées, les travaux sur de la reconnaissance des entités nommées sont une opportunité pour la valorisation de cette langue dans la société de l'information.

C'est dans cette optique que se situe le travail que nous présentons dont le but est de détecter et extraire dans des textes amazighs les entités nommées pertinentes et ceci en développant le premier système de reconnaissance d'entités nommées (RENAM) qui servira à des applications plus spécifiques. Notre système est fondé sur une approche symbolique où l'extraction s'effectue en se basant sur un ensemble de lexiques de noms et des règles construites manuellement en exploitant l'outil d'extraction des entités nommées disponible sous la plateforme GATE[1].

Notre article est structuré comme suit : la première section présente un état de l'art des approches d'extraction des entités nommées, la deuxième examine les difficultés entravant l'extraction des entités nommées en amazighe, la troisième aborde les particularités de la plateforme choisie ainsi que la méthodologie utilisée, tandis que la discussion

[1] http ://gate.ac.uk

des résultats obtenus est l'objet d'étude de la quatrième section. Finalement, la dernière section est consacrée à la conclusion et perspectives.

2 Approches d'extraction d'entités nommées

La REN constitue un champ de recherche très actif. Différentes approches existent, l'extraction fondée sur des démarches linguistiques ou encore nommées symboliques, approches statistiques ou à base d'apprentissage, avec une apparition d'une approche hybride. La première approche s'appuie sur l'utilisation de grammaires formelles construites à la main. Elle se fonde sur la description des EN grâce à des règles qui exploitent des marqueurs lexicaux, des dictionnaires de noms propres et parfois un étiquetage syntaxique. Les marqueurs lexicaux, il s'agit d'indices qui encadrent, l'entité nommée et qui permettent souvent de dévoiler sa présence. D'autre part, les dictionnaires de noms propres groupent généralement une liste des noms et des prénoms les plus fréquents, des noms de localisations, et parfois des noms d'organisations. Ces dictionnaires sont fréquemment utilisés dans les systèmes à base de règles comme ils peuvent avoir recours aux systèmes à base d'apprentissage. Les ressources mentionnées renforcent l'élaboration des règles et des patrons linguistiques qui spécifient les contextes d'apparition de telle entité. Pour les langues peu dotées, cette approche semble la plus adéquate. Dans le cadre des approches linguistiques, nous citons quelques systèmes de REN arabes à savoir : les travaux de Shaalan et Raza (Shaalan et Raza, 2008) qui ont développé leur système NERA permettant d'extraire dix types d'EN. Ce système repose sur l'utilisation d'un ensemble de dictionnaires d'EN et sur une grammaire sous forme d'expressions régulières pour la reconnaissance des EN. Le meilleur taux F-mesure acquis par ce système est 98.6%. Suivant presque le même principe, les travaux de Zaghouani (Zaghouani et al., 2010) ont présenté un module de repérage des EN à base de règles pour la langue arabe, la seule différence c'est qu'ils ont procédé à une première étape de prétraitement lexicale qui prépare le texte pour son analyse linguistique, ce module a été évalué sur un corpus de presse. Dans le même contexte des langues peu dotées nous évoquons le premier système élaboré pour la langue Iban (Malaisie) par Soo-Fong Yong, Bali Ranaivo-Malançon et Alvin Yeo Wee (Soo-Fong Yong, Bali Ranaivo-Malançon et Alvin Yeo Wee, 2011), il consiste à employer des listes des noms propres et d'un ensemble de règles rédigées manuellement pour permettre l'extraction des entités nommées Iban, ils ont obtenu un F-mesure qui est égal à 76,4%. La seconde démarche fait usage de techniques statistiques ou encore dites à base d'apprentissage pour apprendre des spécifiés sur de larges corpus de textes où les entités-cibles ont été auparavant étiquetées nommés ainsi corpus d'apprentissage, et par la suite adapter un algorithme d'apprentissage qui va permettre d'élaborer automatiquement une base de connaissances à l'aide de plusieurs modèles numériques (CRF, SVM, HMM …). Cette méthode a été envisagée pour avoir une certaine intelligence lors de la prise des décisions, ce sont principalement certains paramètres qui peuvent être manipulés dans le but d'améliorer les résultats du système, ce qui n'est pas le cas pour les approches symboliques qui n'appliquent que les règles préalablement injectées. Benajiba et al (Benajiba et al., 2009) ont conçu une technique d'apprentissage SVM pour leur système de REN en se servant d'un ensemble de particularités de la langue arabe. Ce système a produit une F-mesure de 82.71%. Pour les langues peu doté comme le Telugu P Srikanth et Kavi Narayana Murthy (P Srikanth & Kavi Narayana Murthy 2008) ont utilisé la technique CRF pour extraire les entités nommées et ils ont obtenu un score de f-mesure qui est égal à 92%, pour le Bengali en utilisant le SVM, Asif Ekbal et Sivaji Bandyopadhyay (Asif Ekbal & Sivaji Bandyopadhyay 2008) ont réussi à avoir un résultat dont la valeur de f-mesure est de 91,8%. Les deux approches qu'on a cité auparavant ont fait l'apparition d'une troisième approche qui représente une combinaison de ses antécédents, elle utilise des règles écrites manuellement mais construisent aussi une partie de ses règles en se basant sur des informations syntaxiques et des informations sur le discours extraites de données d'apprentissage grâce à des algorithmes d'apprentissage, des arbres de décisions. Abuleil (Abuleil, 2006) a adopté une approche hybride pour l'extraction des entités en arabes, en tirant profit des approches symboliques et d'apprentissage.

3 Difficultés entravant l'Extraction d'Entités Nommées Amazighes

La langue Amazighe est composée de 27 consonnes, 2 semi-consonnes, 3 voyelles pleines et une voyelle neutre. Elle présente une morphologie riche et complexe. Les mots peuvent être classés en trois catégories morphosyntaxiques: Nom, Verbe et Particules[2] (Boukhris et al., 2008). La structure basique de la phrase simple est : Sujet – verbe – complément. Cependant, les travaux liés à la REN de la langue amazighe sont encore à l'état naissant en raison de:

[2] Elles sont un ensemble de mots Amazighs qui ne sont ni des noms, ni des verbes, et jouent un rôle d'indicateurs grammaticaux au sein d'une phrase. Cet ensemble est constitué de plusieurs éléments à savoir: les particules d'aspect, d'orientation et de négation; les pronoms; les adverbes; les prépositions; les subordonnants et les conjonctions.

– L'absence de la distinction majuscule/minuscule : c'est un obstacle majeur pour la langue amazighe. En fait, la REN pour certaines langues comme les langues indo-européennes se base principalement sur la présence des lettres majuscules qui est un indicateur très utile pour identifier les noms propres dans les langues utilisant l'alphabet latin. Il n'y a pas de majuscule, ni au début ni à l'initiale des noms propres amazighe.
– L'amazighe se caractérise par le manque de ressources dictionnairiques, de répertoires toponymiques, de ressources langagières et outils du TAL, à savoir les étiqueteurs, les analyseurs morphologiques.
– Il est un fait que la langue amazighe est très agglutinative ayant une morphologie dérivationnelle et flexionnelle assez complexe.
– Similairement à d'autres langues naturelles, l'amazighe présente des incertitudes au niveau des classes grammaticales. En effet, la même forme convient à nombreuses catégories grammaticales, cela dépend du contexte dans la phrase. Par exemple, ⵉⵍⵍⵉ [illi] peut être considéré comme verbe à l'accompli positif, il signifie «il existe», ou comme nom de parenté «ma fille».
Les noms propres au niveau de la langue amazighe sont extrêmement nombreux, ont de nombreuses variantes ainsi ils sont difficiles à détecter sans la présence d'un lexique.

4 Système de Reconnaissance d'Entités Nommées Amazighes (RENAM)

La tâche de REN amazighes apparaît fondamentale pour diverses applications participant à l'analyse du contenu des textes amazighes. Dans cette contribution, nous nous intéressons à développer un système d'analyse de textes amazighes permettant le repérage et la catégorisation des entités nommées en fonction de types sémantiques prédéfinis. Pour respecter les formats d'annotation standard, nous avons adopté le standard d'étiquetage proposé par la conférence MUC (Message Understanding *Conference*) pour les entités nommées. Par conséquent les EN appartiennent à trois classes majeures à savoir la catégorie **ENAMEX** qui comprend les types « personne », « organisation », et « localisation », la catégorie **TIMEX** qui inclut les types « date », « heure », et finalement la catégorie **NUMEX** qui comprend « montants financiers », « pourcentages » tout en exploitant la plate-forme GATE.

4.1 Plateforme GATE

La plateforme d'ingénierie textuelle **GATE** (General Architecture for Text Engineering) (Cunningham et al. 2002) est une infrastructure de développement de traitement du langage humain développée par l'Université de Sheffield et est exploitée dans une vaste variété de recherche et de projets de développement incluant l'extraction de connaissances pour l'anglais, l'espagnol, le chinois, l'arabe, le français, l'allemand, l'hindi, le cebuano, le roumain, le russe. Nous avons choisi cet environnement car il permet d'implémenter l'architecture souhaitée et il peut être utilisé pour l'exploitation des traitements linguistiques dans diverses applications. Le module d'extraction proposé dans GATE est réalisé selon une approche symbolique basé sur le formalisme **JAPE** (Java Annotation Patterns Engine) permettant de définir les contextes d'apparition des unités à extraire dans le but de les repérer et les annoter.

4.2 Architecture du système

Etant donné la non-disponibilité d'un large corpus pour la langue amazighe, nous avons créé le notre. Notre système d'extraction d'entités nommées est un système de détection et de typage des entités d'intérêt dans un texte. Notre but se limite à l'extraction des entités de type «Personne», «Organisation», et «Localisation», dans des textes écrits en amazighe, en utilisant une approche symbolique en créant manuellement des règles pour extraire les informations qui nous intéressent.

Notre système RENAM se compose de deux grandes phases successives (Figure 1): la première consiste à la préparation des données et la deuxième consiste à la reconnaissance des entités nommées et classification de ces dernières dans des catégories sémantiques convenables (personne, localisation, organisation).

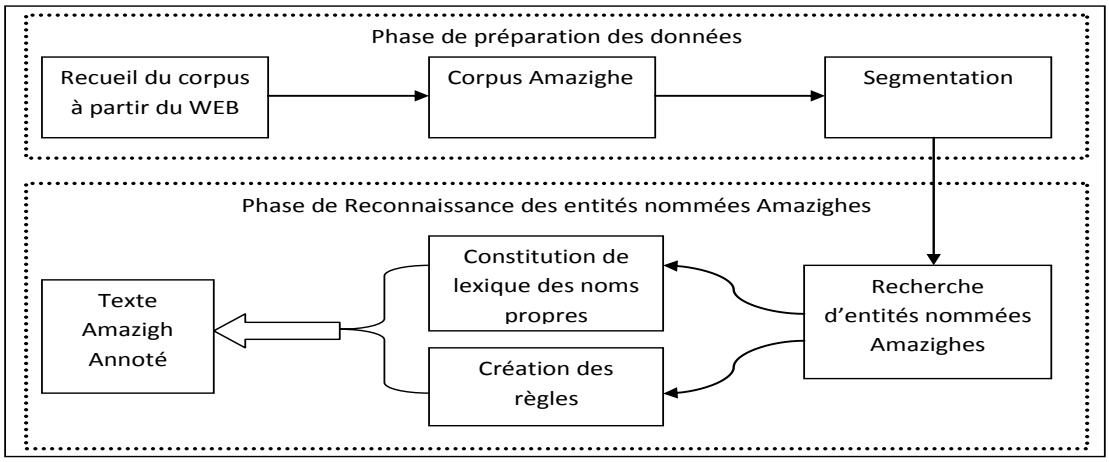

FIGURE 1: Architecture du Système d'Extraction des Entités Nommées Amazighes

– Phase de préparation des données

1. Recueil du corpus

La construction d'un système d'extraction d'entités nommées exige, d'abord, de rassembler un nombre suffisant de textes qui serviront non seulement de corpus d'observation (pour constituer les règles) mais également de corpus de test. Le contexte de cette contribution nous a naturellement guidé vers la collecte de textes amazighes journalistiques à partir du site web « mapamazighe[3]». Le corpus contient toute l'actualité sur les activités royales de SM le Roi Mohammed VI, on dispose actuellement de 200 articles écrits en amazighe, que nous avons convertis du format « html » au format « txt », cumulant un total de 50997 mots, nous signalons aussi qu'outre ce corpus inclut 4866 chiffres.

2. Segmentation du corpus

Dans cette phase, nous séparons le texte amazigh en des phrases, nous avons en total 1214 phrases, et ensuite nous séparons les phrases en des mots. Dans le texte, les phrases sont fragmentées selon la présence des ponctuations et le retour à la ligne, comme dans le cas dans les langues française et anglaise. Ces séparateurs définissent les modèles de caractères marquant la fin des phrases.

– Phase d'extraction des entités nommées

3. Constitution de lexique des noms propres

Dans le but de pouvoir reconnaître automatiquement les entités nommées de type personne, organisations et localisation, nous avons commencé par l'élaboration manuelle d'un lexique provenant de plusieurs sites Web.
Pour les entités nommées de type « personne», nous avons pu construire une liste d'environ 1650 entrées de noms amazighes et noms transcrits en amazighe[4]. Pour les entités nommées de type « localisation », nous nous sommes inspirés de la classification faite par (Piton et Maurel, 2004) qui considère comme type « localisation » ou toponyme : les pays, villes, fleuves, montagnes, océans et mers. Ainsi, nous avons élaboré un lexique à partir de wikipedia (nous avons procédé à la transliteration des entités françaises extraites à partir de wikipédia vers l'amazighe), et le site officiel de l'IRCAM qui contient 1970 entités de type « localisation ». A l'instar de la procédure d'élaboration de lexiques des deux entités nommées précédentes, l'identification des entités nommées de type organisation a commencé par le développement d'un lexique qui contient 120 noms d'organisation, à partir de notre corpus qui contient un nombre assez important des noms d'organisations, et les sites web officiels des organisations en question.
Nous avons aussi élaboré un lexique des marqueurs lexicaux de type localisation avec 55 entrées, de type personne avec 84 entrées et de type organisation avec 70 entrées.

[3] http://www.mapamazighe.ma/am/

[4] www.dsic.upv.es/%7Eybenajiba/resources/ANERGazet.zip

4. Développement des règles linguistiques

Pour la construction de nos règles linguistiques pour le traitement de l'amazighe, nous nous sommes inspirés des travaux de (McDonald, 1996) dans le domaine qui consiste à déterminer les indices d'apparition des entités à extraire. Ainsi, ces derniers font partie intégrante de l'entité nommée comme le prénom pour une entité de type «Personne » ou correspondent au contexte textuel de l'entité pouvant indiquer son type. La découverte des marqueurs lexicaux s'est faite par observation du corpus décrit précédemment.

Règles de type « Personne » :

Les règles pour la détermination de l'entité personne fait usage à la présence :

- Des civilités à titre d'exemple : ⵎⴰⵙⵙ [mass] « Monsieur », ⵜⴰⵎⵎⴰ [mast] « Mme », ⵜⴰⵎⴰⵙⵙⴰ [massa] « Mme », ⵍⴰⵍⵍⴰ [lalla] « Mme », ⵎⵓⵍⴰⵢ [moulay] « My » …
- Les titres de toute sortes tels que :
  - Titres politiques (40 entrées) : ⴰⵎⴰⵡⴰⵙ [amawas] « ministre »
  - Titre honorifiques (14 entrées) : ⴱⴰⴱ ⵏ ⵡⴰⵏⵏⵓⵔ [bab n waddur] « Sa Majesté »
  - Titre religieux (8 entrées) : ⵛⵉⵅ [chikh] « Cheikh »
- Les métiers ou professions (… entrées) : ⵍⴽⵓⵍⵓⵏⵉⵍ [lkulunil] « le colonel »

Pour définir le lien de parenté, chaque culture a son propre système de nommage, par exemple ⵎⵓⵃⵎⵎⴷ ⴱⵏ ⵣⴰⵢⴷ [muhmmd bn zayd] « Mohamed Ibn Zayd », le modèle suivant indique que cette séquence est un nom de personne :

<NomPropre> <ⴱⵏ (bn, fils) > <NomPropre> ➔ Nom= « entité personne ».

Règles de type « Localisation » :

Les règles linguistiques pour la reconnaissance des entités de *type « localisation »* sont écrites en se basant sur des prépositions (15 entrées) qui accompagnent toujours les noms de lieux comme par exemple ⴷⴰⵔ [dar] « en direction de », ⵖⵔ [ghr] « vers », ⵊⴰⵔ…ⴷ… [jar…d…] « entre… et… »…

En outre, nous avons construit une liste de marqueurs lexicaux comme par exemple, ⵜⵎⴷⵉⵏⵜ [tmdint] « ville ». Ces deniers sont exploités pour la description des règles de reconnaissance des entités de *type « localisation ».*

Par exemple, ⴰⵙⵉⴼ ⵏ ⵙⴱⵓ [Asif n sbu] « rivière de sbou », ce qui signifie que tout mot qui vient après un préfixe de localisation (dans ce cas ⴰⵙⵉⴼ [asif] « rivière ») sera annoté comme étant un nom de lieu.

Règles de type « Organisation » :

Nous avons élaboré une liste des marqueurs lexicaux comme par exemple ⵜⴰⵎⴰⵡⴰⵙⵜ [amawast] « Ministère » pour la description des règles de reconnaissance des entités de *type « organisation »*

Tableau récapitulatif de règles construites :

Entités nommées	Exemple	Nombre de règles
Personne	ⵎⴰⵙⵙ(Monsieur), ⵜⴰⵎⵎⴰ(Madame), ⵙⴰⵙ ⵏ ⵡⴰⵏⵏⵓⵔ (Sa majesté)…	10
Organisation	ⵜⵉⵎⵙⵙⵏ(Ecole), ⵜⴰⵎⴰⵡⴰⵙⵜ(Ministère) …	4
Localisation	ⴰⵙⵉⴼ(Fleuve), ⵜⵎⴷⵉⵏⵜ(Ville) …	5

TABLE 1 : Tableau récapitulatif

5 Evaluation des résultats

Une dernière étape de cette contribution est l'évaluation des résultats de notre outil d'extraction des entités nommées amazighes. Ainsi, nous présentons, dans cette section, les différentes étapes de notre évaluation et les conclusions que nous avons pu faire en vu des résultats obtenus.

5.1 Protocole expérimental

L'évaluation de notre système a, ainsi, permis d'établir plusieurs choix visant le type d'évaluation à mettre en place. Deux solutions se sont alors présentées : exploiter les données d'une campagne d'évaluation existante ou créer notre propre système d'évaluation. Le premier cas n'étant pas possible puisqu'il n'existe pas un corpus de la langue amazighe, où les entités nommées ont été préalablement annotées. En conséquence, nous avons envisagé la deuxième solution, qui est le développement de notre système d'évaluation.

5.2 Analyse des résultats

Dans notre évaluation, nous avons choisi les métriques qui nous permettront d'évaluer les résultats de nos travaux et ainsi mesurer les performances du système proposé.

La précision (mesure de qualité), le rappel (mesure de quantité) et la F-mesure (synthèse de Rappel et de précision) (*Van Rijsbergen, 1979*), ont été choisis pour leur exploitation fréquente dans le domaine du TAL.

$$\text{Rappel} = \frac{nombre\ d'entités\ correctement\ étiquetées}{nombre\ d'entités}$$

$$\text{Précision} = \frac{nombre\ d'entités\ correctement\ étiquetées}{nombre\ d'entités\ étiquetées}$$

$$\text{F-Mesure} = \frac{2 \times Précision \times Rappel}{Précision + Rappel}$$

Nous avons effectué une évaluation sur notre corpus « Activités Royales » que nous avons déjà décrit, et cela a donné les résultats suivants :

Entités nommées	Précision	Rappel	F-mesure
Personne	64%	63%	64%
Organisation	27%	71%	40%
Localisation	82%	81%	82%

TABLE 1 : Evaluation de notre système de reconnaissance des entités nommées

D'après la Table 2, les résultats que nous avons obtenus sont plus ou moins satisfaisants et encourageants. Cependant notre outil fonctionne légèrement moins bien pour l'entité nommée « organisation ».

Les principales causes de cette déficience sont :

- On a remarqué que les noms d'organisation sont très peu utilisés par rapport aux autres entités nommées.

- Les entités nommées complexes ne sont pas correctement extraites. A titre d'exemple ; le mot déclencheur ⵜⴰⵎⴰⵡⴰⵙⵜ [tamawast] « Ministère » qui indique que la séquence qui va suivre représente une entité nommée de type organisation, néanmoins l'absence des informations morphologiques nécessaires rend la tâche de délimitation de l'entité en question plus difficile. Par exemple : ⵜⴰⵎⴰⵡⴰⵙⵜ ⵏ ⵙⴳⵎⵉ ⴰⵏⴰⵎⵓⵔ [tamawast n usgmi anamur] « Ministère de l'éducation nationale ». Notre outil ne va reconnaitre que l'entité « ⵜⴰⵎⴰⵡⴰⵙⵜ ⵏ ⵙⴳⵎⵉ », parce qu'il n y a pas de critère d'arrêt selon chaque entité.

– La prise en compte des variantes orthographiques des noms propres transcrits en l'absence de conventions pour leurs écritures (notamment pour les noms de lieux). En amazighe, la translittération et la transcription des noms propres étrangers n'obéissent pas à des règles d'écritures par exemple ⴰⴱⵓ ⴷⴰⴱⵉ [abu Dabi] « Abu Dhabi » ou ⴰⴱⵓⴷⴰⴱⵉ [abudabi] « Abudhabi »).

6 Conclusion et perspectives

L'objectif principal de cette contribution est de reconnaître les entités nommées de types (personne, localisations, organisations,) dans les textes amazighes. De ce fait, nous avons élaboré dans un premier temps, un système de reconnaissance d'entités nommées pour l'amazighe, fondé sur une approche symbolique qui se base sur un ensemble de règles linguistiques construites manuellement et sur l'élaboration des lexiques de noms en se servant de la plateforme GATE. La valeur de F-mesure obtenue par notre méthode est assez encourageante malgré la présence des problèmes qu'on a décrit auparavant. Dans un futur immédiat, nous envisageons améliorer la qualité des règles d'extraction d'entités nommées que nous avons établies, étendre et enrichir la structure de nos lexiques de noms afin de couvrir le maximum des entités nommées et pour aboutir à un taux de reconnaissance plus élevé, étendre la taille de notre corpus, ensuite passer au traitement des deux autres catégories TIMEX (Date…) et NUMEX (pourcentage, …).

Références

BENAJIBA Y., ROSSO P. (2008). Arabic Named Entity Recognition using Conditional Random Fields. *In Proceedings of Workshop on HLT and NLP within the Arabic World*, *LREC*.

BICKEL M., MILLER S., SCHWARTZ R., WEISCHEDEL R. (1997). "Nymble: a high-performance learning name-finder". *In Proceedings of the ANLP 97.*

BORTHWICK A., STERLING J., AGICHTEIN E., GRISHMAN R. (1998). "Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition". *In Proceedings of the WVLC 98.*

BOUKHRIS F., BOUMALK A., ELMOUJAHID E., SOUIFI H. (2008). La nouvelle grammaire de l'amazighe. Rabat, Maroc: IRCAM

CHAKER S. (1984). Textes en linguistique berbère - introduction au domaine berbère, *éditions du CNRS*, 232-242.

COHEN D. (2007). Chamito-sémitiques (langues). *In Encyclopædia Universalis*.

CUNNINGHAM H. (1999). Information Extraction: a User Guide (revised version), *Research Memorandum*, Department of Computer Science, University of Sheffield

CUNNINGHAM H., MAYNARD D., BONTCHEVA K., TABLAN V., URSU C. (2002). The GATE User Guide.

EHRMANN M. (2008). Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation. *PhD thesis*, Université Paris 7

FONG Y., BALI R., WEE A. (2011). NERSIL - the Named-Entity Recognition System for Iban Language. *PACLIC 2011:* 549-558

GAHBICHE B S., BONNEAU H., MAYNARD D., LAVERGNE T, YVON F. (2012), Repérage des entités nommées pour l'arabe, *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, 487–494*

MAYNARD D. (*2011*). Developing Language Processing Components with GATE, Version 6, *http://gate.ac.uk/sale/tao/tao.pdf.*

MCDONALD D. (1996). Internal and external evidence in the identification and semantic categorization of proper names. *In B. Boguraev and J. Pustejovsky, editors, Corpus Processing for Lexical Acquisition,* 21–39.

OANA H., BONTCHEVA K., MAYNARD D., TABLAN V., CUNNINGHAM H. (2003). Named entity Recognition in Romanian using Gate. *RANLP 2003 Workshop on information Extraction for slavonic and other central and eastern European languages*, Borovets : Bulgaria

GRISHMAN R. (1995). The NYU system for MUC-6 or Where's the Syntax. *In the proceedings of Sixth Message Understanding Conference (MUC-6),* 167-195

SHAALAN K., RAZA H. (2009). NERA : Named entity recognition for arabic. *Journal OF the American Society for Information Science and Technology*, 1652–1663.

TAKEUCHI K., COLLIER N. (2002). Use of support vector machines in extended named entity. *In*: *Proc*. *CoNLL-2002*

TJONG K. S., DEMEULDER F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, *in: Proceedings of the 7th Conference on Natural Language Learning*, Edmonton, Canada, 142–147.

WAKAO T., GAIZAUSKAS R., WILKS Y. (1996). Evaluation of an algorithm for the recognition and classification of proper names. *In Proceedings of COLING-96*.

ZAGHOUANI W., POULIQUEN B., EBRAHIM M., STEINBERGER R. (2010). Adapting a resource-light highly multilingual named entity recognition system to arabic. *Proceedings of the Seventh conference on International Language Resources and Evaluation* (*LREC'10*), 563–567.