



HAL
open science

La détermination des attributs sémantiques internes à la structure syntaxico-sémantique des entités nommées en vue de leur extraction automatique

Omar Asbayou

► To cite this version:

Omar Asbayou. La détermination des attributs sémantiques internes à la structure syntaxico-sémantique des entités nommées en vue de leur extraction automatique. CIDE 20, ENSIB, Nov 2017, LYON (ENS et Lyon 3), France. hal-04683360

HAL Id: hal-04683360

<https://hal.science/hal-04683360v1>

Submitted on 2 Sep 2024

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

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

La détermination des attributs internes à la structure syntaxico-sémantique des entités nommées en arabe en vue de leur extraction

The specification of arabic named entity internal attributes for their automatic extraction

Omar ASBAYOU,

ICAR (UMR 5191), Université Lumière Lyon 2
omar.asbayou@univ-lyon2.fr

Résumé. Les classes des EN (entités nommées) ont des attributs sémantiques. Ces attributs, qui sont en nombre fini et auquel sont associés des valeurs représentées par des entités linguistiques bien déterminées, constituent des éléments essentiels dans la construction des structures syntaxico-sémantiques. La détermination de ces derniers constituent une étape indispensable pour notre système à base de règles d'extraction et de classification automatique des EN en arabe. Dans cet article, nous allons nous limiter d'exposer les différents attributs sémantiques des classes PERSONNE, ORGANISATION, LIEU, EVENEMENT et ARTIFACT- LOI. Ces attributs nous ont été non seulement essentiels mais aussi indispensables dans la construction de notre système à base de règles de reconnaissance et de classification des EN en arabe dont les résultats (corpus annoté) sont exploités dans le traitement du contenu textuel et plus particulièrement dans les deux domaines de la recherche d'informations et l'extraction d'informations.

Mots-clés. Entités nommées, extraction d'informations, recherche d'informations, attributs sémantiques.

Abstract. NE (named entities) have semantic attributes. These attributes, which are in a finite number and to which values represented by linguistic entites are associated, constitute essential elements in syntactico-semantic structure construction. Their determination is an indispensible step for our rule-based Arabic NE extraction and classification system. We will limit ourselves in this paper to exposing the different semantic attributes of NE classes PERSON, ORGANISATION, LOCATION, EVENT, and ARTIFACT-LAW. The results of our NE extraction and classification system (annotated corpus) are exploited in textual contents processing, partcularly in information retrieval and information extraction.

Keywords. Named entities, information extraction, information retrieval, semantic, attributes.

1 Introduction

Ce travail s'inscrit dans un cadre général de traitement automatique des langues, mais plus particulièrement dans la continuité des travaux réalisés au niveau de l'analyse morphosyntaxique par la conception et la réalisation des bases des données lexicales SAMIA (Synthèse et Analyse Morphosyntaxiques Informatisées de l'Arabe) et ensuite DIINAR (Dictionnaire Informatisé de l'Arabe) avec l'ensemble de résultats de recherches qui en découlent. Nous devons citer dans ce cadre J. Dichy (1984,1990, 1997, 2000), Hassoun (1987) Dichy et Hassoun (2005).

Notre tâche vise, en mettant en œuvre les résultats de ce traitement morphosyntaxique basé sur DIINAR, à l'enrichissement lexical par des EN simples et complexes, et qui veut établir la transition de l'analyse morphosyntaxique vers l'analyse syntaxique, et syntactico-sémantique dans une visée plus générale de l'analyse du contenu textuel.

Les EN englobent les noms propres, les descriptions définies, les entités temporelles et les entités numériques. Ce sont des unités linguistiques à forte valeur informative. Leur reconnaissance contribue à une représentation sémantique du contenu textuel et constituent des éléments d'indexation importants et pertinents pour la fouille de texte.

Notre système d'extraction et de classification des EN en arabe basé sur des patrons syntactico-sémantiques entend son intégration dans des applications appropriées de traitement automatique de des langues, notamment, à l'extraction et la recherche d'informations.

La spécification des attributs sémantiques des classes, qui est une étape essentielle dans le processus de construction de nous règles, rend notre système adaptable et utile aux applications de recherche et d'extraction d'informations ; cette étape prenant en considération l'aspect syntactico-sémantique et sémantico-conceptuel permet d'attribuer aux EN des annotations fines en sortie du traitement linguistique. Ces résultats servent de base pour une indexation particulière basée sur les classifications fournies par les annotations sémantiques (un corpus annoté en sortie) pour améliorer la précision de la recherche d'informations et de l'extraction d'informations.

2 Les attributs de la classe « PERSONNE »

La classe "PERSONNE » a les attributs suivants qui peuvent se manifester par des constituants d'EN.

- **Nom propre : nom et prénom** : Pour les personnes, le nom est un pur nom propre comme ديكارت رينه (René Descart) ou تشومسكي نعوم (Noam Chomsky) composé d'un prénom (dans la liste des noms propres de personnes), souvent partagé par un ensemble d'individus, et d'un nom de famille, qui représente l'entité distinctive principale entre les noms de personnes.
- **Fonction** : La fonction de la classe « PERSONNE » se traduit par les déclencheurs (ou sous-classes) :
 - *entité simple* : ex. الفيلسوف (le philosophe)
 - *entité complexe* (description définie) : ex. عالم اللسانيات (le linguiste)
- **Domaine** : Le domaine est montré par les deux entités principales de l'entité nommée :

- *le déclencheur spécifique* : il dénote un domaine d'intérêt. Par exemple :
 - a) العالم (le savant) / الفيلسوف (le philosophe) => Science
 - b) وزير (ministre) => Politique
 - c) جنرال (Général) => Sécurité
- *l'extension* : dans le cas où le déclencheur est générique (ne spécifie pas le domaine), on se base sur l'extension pour extraire le domaine. Par exemple :
 - a) رئيس كلية الآداب والعلوم الإنسانية (le président de la **faculté des lettres et des sciences humaines**) => Science
 - b) الكاتب العام لحزب العدالة والتنمية (le secrétaire général du **parti de la justice et de développement**) => Politique

L'extension est dans la plupart des cas une EN « ORGANISATION » come dans c) et d).

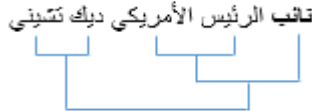
- **Personne** : Une personne peut avoir une autre personne comme attribut : une relation professionnelle ou familiale entre deux personnes. Les entités qui correspondent à cet attribut sont généralement des éléments de base dont l'expansion d'annexion est l'entité « PERSONNE ». Regardons ces exemples :

- a) نجل الزعيم الليبي (Le **fils** du dirigeant libyen) => attribut familial
- b) نائب الرئيس الأمريكي (le **vice-président** américain) => attribut professionnel

Les noms propres, qui peuvent survenir dans cette structure (généralement en position final), doivent donc être des attributs des attributs ; c'est-à-dire, l'attribut de cette entité occupant la position de la tête du syntagme.

- c) نائب الرئيس الأمريكي **ديك تشيني** (le **vice-président** américain **Dick Cheney**)

L'analyse en constituants immédiats de cette EN se fera comme suit :

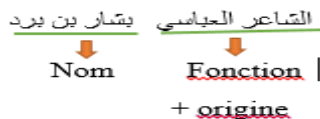


Cette analyses permet donc d'assigner le nom propre **ديك تشيني** (*Dick Cheney*) à l'entité **نائب الرئيس الأمريكي** (le vice-président américain) et non pas à **الرئيس الأمريكي** (le président américain)

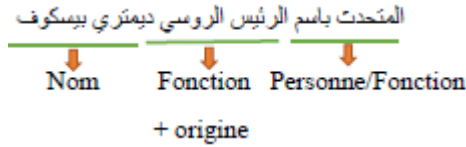
- **Origine** : Chaque personne à une origine exprimé par un nom propre de lieu ou par sa nationalité (nom commun adjectif). Par exemple :
- **Nationalité** : الأديب الفرنسي (l'écrivain **français**)
 - **Pays** : رئيس فرنسا (le président de la **France**)

Les attributs mentionnés ci-dessus peuvent se retrouver dans une seule entité nommée. Regardons ces exemples :

Eg.1 الشاعر العباسي بشار بن برد (le poète abbasside, Bachchâr Ibn Burd)



Eg.2 المتحدث باسم الرئيس الروسي ديمتري بيسكوف (le porte-parole du président russe Dimitri Peskov)



Nous remarquons ici que les entités sont composées d'un ensemble d'attributs de sa classe (« PERSONNE »).

- pour la sous-classe « PERSONNE » de الرئيس الروسي (le président russe) nous avons l'attribut "personne-fonction" réalisé par المتحدث باسم (le porte parole).
- Attribut « fonction » : الشاعر (le poète) dans eg.1, et l'ensemble de la description définie المتحدث باسم الرئيس الروسي (le porte-parole du président russe) dans eg.2.
- Attribut « origine » : العباسي (Abbasside) dans e.g1, et الروسي (russe) dans e.g2. Cet attribut est réalisé par les éléments de deux types de listes: liste de noms de nationalités et liste de noms de pays.
- Attribut "nom": بشار بن برد (Bachchâr Ibn Burd) dans eg.1, et ديمتري بيسكوف (Dimitri Peskov) dans eg.2. Les noms propres de personnes sont soit dans des répertoires ou bien reconnus par des règles d'extraction.

3 Les attributs de la classe « LIEU » :

Les lieux ont les attributs suivant :

- **Nom (pur nom propre) :**
 - a) شارع شارل دوكلو (rue Charles De Gaule)
 - b) مدينة سان فرانسيسكو / باريس (ville de San Francisco/Paris)
 - c) جبال الألب (les Alpes).
- **Bâtiment :** Cet attribut est associé aux entités avec un local comme les organisations. Il est exprimé par le déclencheur :
 - a) مقر مجلس الأمن (le siège du Conseil de sécurité)
 - b) بناية البنتاغون (le bâtiment du Pentagone)
- **Fonction :** Elle correspond à la fonction de l'organisation associée, par une relation de métonymie, à ce lieu. Cet attribut est exprimé par le SP (syntagme prépositionnel) de l'extension exprimant le but comme dans مركز دبي للإحصاء (le Centre de Dubaï pour les statistiques).

4 Les attributs de la classe « ORGANISATION » :

En traitant les attributs des EN, on peut constater que la classe « ORGANISATION » occupe la position centrale dans l'extraction des EN. Les attributs des EN « ORGANISATION » peuvent être déterminés par des entités avant ou après leurs déclencheurs.

Avant de traiter les attributs représentés par des constituants qui s'affichent avant ou après le déclencheur, le premier attribut de la classe « ORGANISATION » est :

- **Nom** : L'attribut "nom" sont deux types de noms propres :
 - *Pur nom propre* : comme فولكسفاكن (Volkswagen).
 - *EN descriptive* : comme وزارة التعليم العالي (le Ministère de l'éducation supérieur).

L'attribut "pur nom propre" peut également être classé parmi les attributs après le déclencheur de type preuve externe comme dans شركة فولكسفاكن (société Volkswagen) (voir ci-après).

A) **Attributs avant le déclencheur d'«ORGANISATION» :**

Nous entendons par "attribut avant le déclencheur", l'attribut exprimé par une entité (un autre déclencheur) précédant le déclencheur de l'EN dont il est attribut.

Pour la classe « ORGANISATION », les attributs avant le déclencheur sont :

- **Personne** : Le type de personnes qui nous intéresse ici est les sous-classes dénotant des fonctions de personnes dans une organisation (descriptions définies), en opposition aux noms propres. Regardons les exemples suivants :

a) رئيس المجلس الوطني لحقوق الإنسان (le président du Conseil national des droits de l'Homme).

b) عميد كلية الآداب والعلوم الإنسانية (le doyen de la Faculté des Letters et des Sciences Humaines)

c) المدير العام للشركة العامة (le président général la Société Générale)

Nous avons donc ici trois entités distinctes :

- Fonction de « PERSONNE » => attribut d'« ORGANISATION »
- EN « ORGANISATION » => extension par subordination par rapport à l'entité exprimant cet attribut « personne ».
- EN« PERSONNE » => l'ensemble de l'entité qui, comme nous avons vu ci-dessus, a à son tour son attribut « nom propre ».

- **Lieu** : La classe "ORGANISATION" a aussi un lieu en tant qu'attribut. Les lieux des organisations sont exprimés pas une liste très limitée de déclencheurs de lieu comme مقر (siège), بناية (bâtiment), مبنى (bâtiment). Par exemple :

a) مبنى مجلس الأمن (Le bâtiment du Conseil de sécurité)

b) مقر وزارة الداخلية (le siège du Ministère de l'intérieur)

c) بناية البنتاغون (le bâtiment du Pentagone)

Nous insistons donc sur le fait que les déclencheurs représentant les attributs des EN auxquelles ils sont attachés en position initiale constituent de nouvelles EN prenant à chaque fois la classe de ces derniers. Cela nous permet d'élargir notre classification selon les besoins en rajoutant, dans cette position initiale dans nos règles d'extraction (les patrons syntaxico-sémantiques), de nouvelles classes d'attributs pour extraire de nouvelles classes d'EN.

- **Loi** : l'attribut loi figure aussi parmi les attributs de la classe ORGANISATION avant le déclencheur. Ce qui ne permet d'obtenir des EN « LOI ». Par exemple :

a) قرار مجلس الأمن رقم 2216 (la résolution du Conseil de sécurité n° 2216).

b) القانون الأساسي لمجلس الأمن (le règlement du Conseil de sécurité).

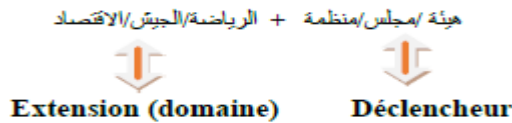
B) Les attributs après le déclencheur d'«ORGANISATION» :

Les attributs après le déclencheur sont exprimés dans l'extension de l'entité « ORGANISATION ». On distingue entre trois attributs :

- **Domaine** : les extensions classées par des annotations sémantiques sont des syntagmes (nominaux et adjectivaux) dénotant chacun un domaine (i.e. champ sémantique). Elles permettent d'affiner la classification des EN.

Ces entités correspondent à des entités morphosyntaxiques et syntaxiques suivants :

- **Noms** : ce sont des mots de domaines (champs sémantiques) : C'est une expansion d'annexion par rapport au déclencheur : الاقتصاد, السياسة, الجيش etc. (Sport, armée, économie etc.) dans :



- **Adjectifs** : cette extension correspond à l'expansion d'identification comme الاقتصادي, السياسي, العسكري, الرياضي etc. (sportif, militaire, politique, économique etc.) dans :

الهيئة / المجلس / المنظمة + الاقتصادي / الاقتصادية : السياسي / السياسية : الرياضي / الرياضية



- **Syntagmes nominaux composés** : nous deux types de SN (syntagmes nominaux) :

- ✓ SN (syntagme nominal) avec Expansion d'identification :

- الصناعة التقليدية (l'artisanat) => [Nom de domaine + adjectif] => SN_ Déterminé _expansion d'identification _Économie
 - البحث العلمي (la recherche scientifique) => [Nom + adjectif de domaine] => SN_ Déterminé _expansion d'identification _Science
- Comme le montrent a) et b), la dénotation du domaine peut être évoqué aussi bien par le nom (la tête) dans (a) (domaine économique) que l'adjectif (le modifieur) dans (b) (domaine scientifique).

- ✓ SN avec Expansion d'annexion :

- حقوق الطفل (les droits de l'enfant) => [Nom + Nom de domaine] => SN_ Déterminé _expansion d'identification _Science
- حقوق المرأة والطفل (les droits de femme et de l'enfant) => [Nom + Nom de domaine + Nom de domaine] => SN_ Déterminé _expansion d'identification _Science

- **SP** : ce type de syntagmes joue le rôle de complément. Ils sont généralement marqués par le proclitique prépositionnel de cause « ل / li / ». On peut les décrire par la formule suivante :

- التعليم (de l'éducation) => SP_ Déterminé _Science
- للتعليم الابتدائي (de/pour l'éducation primaire) => SP_ Déterminé _expansion d'identification _Science
- لحقوق الإنسان (des droits de l'Homme) => SP_ Déterminé _expansion d'annexion _Société

- **Fonction** : L'attribut "fonction" de la classe « ORGANISATION » est exprimée dans les extensions soit par un SN ou par un SP. Par exemple :

a) المجلس الأعلى للاستثمار (le Conseil supérieur de l'investissement) => SP_Déterminé_Economie

b) هيئة تنظيم سوق العمل (le Comité de régulation du marché de l'emploi) => SN_Déterminé_Economie

- **Origine** : L'origine d'une organisation peut être trouvée dans l'extension. Elle est exprimée par le nom propre de lieu ou par une nationalité :

- **Nationalité** : الوكالة الأمريكية للتعليم (l'agence américaine de l'éducation) => Nationalité

- **Pays** : هيئة تنظيم سوق العمل في البحرين (le Comité de régulation du marché de l'emploi au Bahreïn) => Nom de pays

5 Les attributs de la classe « EVENEMENT »

Les Attribut de la classe « EVENEMENT » sont :

- **Nom** : Comme c'est le cas pour toutes les classes, « EVENEMENT » a "nom" comme attribut. Par exemple :

الدورة التاسعة والعشرين للألعاب الأولمبية في الصين (la vingt-neuvième édition des jeux olympiques 2008 en Chine).

A partir de cet exemple, on peut illustrer les attributs de la classe « EVENEMENT ».

A) Les attributs après le déclencheur :

- **Nombre** : التاسعة والعشرين (vingt-neuvième)

- **Date** : 2008

Les dates, en tant qu'attribut d'événement, peuvent se transformer en nom d'événement comme الحادي عشر من سبتمبر (le 11 septembre).

- **Lieu** : الصين (Chine)

- **Domaine** : sport

- **Fonction** : للألعاب الأولمبية (des jeux olympiques)

Ces attributs peuvent être exprimés par différents types de syntagmes (nominaux ou prépositionnels) selon le type de déclencheur (simple ou complexe) :

a) مؤتمر دعم وتنمية الاقتصاد المصري (le Congrès de développement de l'économie égyptienne).

b) المؤتمر الوطني للتشغيل (le Congrès national pour l'emploi)

B) Les attributs avant le déclencheur :

- **Personne** : Cet événement doit avoir bien évidemment l'attribut "personne" comme dans :

مدير الدورة التاسعة والعشرين للألعاب الأولمبية في الصين (le directeur de la vingt-neuvième édition des jeux olympiques 2008 en Chine)

Le déclencheur dans a) exprime l'attribut avant le déclencheur de l'entité «EVENEMENT» et crée une nouvelle entité : sous-classe

« PERSONNE_ADMINISTRATIVE » de la classe « PERSONNE ».

On peut avoir, dans le domaine des EN, d'autres attributs selon

l'enrichissement de la classification par des sous-classes

- **Loi** : Cette entité est une sous-classe « LOI_EVENEMENT » de la classe « ARTEFACT ». Comme c'est le cas pour la classe « ORGANISATION » ci-dessus, « EVENEMENT » peut avoir "loi" comme" attribut. . Par exemple, on a rajouté une sous classe « LOI_EVENEMENT ». Regardons cet exemple :

قرار المؤتمر الوطني العام رقم 7 لسنة 2012 (la résolution du Congrès National Général numéro 7 pour l'année 2012).

6 Les attributs de sous-classe « LOI » :

La sous-classe « LOI », qui appartient à la classe « ARTEFACT », présente les attributs suivants :

- **Nom** : Par exemple :
 - a) الفقرة الثانية من المادة 29 من القانون رقم 49 لسنة 1977 (le deuxième paragraphe de l'article 29 de la loi numéro 49 de l'année 1977)
 - b) مرسوم رقم 76-63 المؤرخ في 25 مارس 1976 المتعلق بتأسيس السجل العقاري (décret n° 63-76 du 25 mars 1976 concernant l'établissement du registre immobilier)

A partir de b) ci-dessus, on reconnaît ces attributs après le déclencheur :

- **Nombre** : رقم 76-63 (n° 63-76)
- **Date** : المؤرخ في 25 مارس 1976 (daté le 25 mars 1976)
- **Domaine** : économie (السجل العقاري registre immobilier)
- **Fonction** : المتعلق بتأسيس السجل العقاري (concernant l'établissement du registre immobilier)

7 Les attributs de la sous classe « ŒUVRE-INTELLECTUELLE » :

قراءات معاصرة في تراثنا (Lectures contemporaines de notre patrimoine philosophique), a les attributs suivants :

- **Nom** : قراءات معاصرة في تراثنا الفلسفي (Lectures contemporaines de notre patrimoine philosophique)
- **Auteur** : محمد عابد الجابري (Mohamed Abed Al-Jabri)
- **Date d'édition** : 2006
- **Maison d'édition** : مركز دراسات الوحدة العربية (Le Centre des Etudes de l'Union Arabe)
- **Domaine** : الفلسفة (philosophie)

Ces informations peuvent être obtenues à partir des citations. Comme on peut le constater, les valeurs des attributs appartiennent à des classes et sont donc en relation avec les attributs de ces dernières. Par exemple, محمد عابد الجابري (Mohamed Abed Al-Jabri) est dans la classe « PERSONNE » et il est donc en relation avec les valeurs des attributs *origine* (Maroc), *fonction* (philosophe) etc. qui peuvent être extraits à partir de l'entité nommée محمد عابد الجابري المغربي (le philosophe marocain Mohamed Abed Al-Jabri). Cela aboutie à l'enrichissement d'un réseau de relation conceptuelle de données.

8 Application

La définition des attributs, chacun étant représenté par une sous-classe bien déterminée, joue un rôle très important dans la construction de nos patrons syntactico-sémantiques de reconnaissance et de classification des EN. Les résultats obtenus dans ce domaine sont importants dans le domaine de la recherche et de l'extraction des informations ; ils permettent de faire le filtrage et la classification des informations par les différents attributs prédéterminés. Pour pouvoir réaliser ce filtrage on doit passer par plusieurs étapes de traitement linguistique :

- Traitement morphosyntaxique : il prend en entrée le texte brut et donne, en sortie, un texte annoté morphosyntaxiquement.
- Traitement syntactico-sémantique : il est basé sur des règles syntactico-sémantiques. il prend en entrée les résultats du traitement morphosyntaxique pour extraire et classifier des entités complexes en leurs associant des annotations syntactiques et sémantiques.
 - Les résultats du traitement syntactico-sémantique servent comme base d'indexation pour la recherche d'information, et de remplissage de rubrique pour un système d'extraction d'information.
 - L'affichage peut donc se faire dans des interfaces de moteurs de. Ces interfaces nous donnent les moyens de filtrage par nom propre d'une classe (personne, organisation, lieu, événement etc.), par domaine (scientifique, littéraire, politique, social etc.). Concernant l'extraction d'information, les rubriques sont remplies on se basant sur les annotations sémantiques obtenues par le traitement syntactico-sémantique.

Pour illustration, voici deux exemples montrant le rôle des attributs des EN dans les applications de recherche d'informations et d'extraction d'informations :

A) Recherche d'information :

Regardons ces deux exemples d'images d'une interface de recherche d'informations (moteur de recherche) :

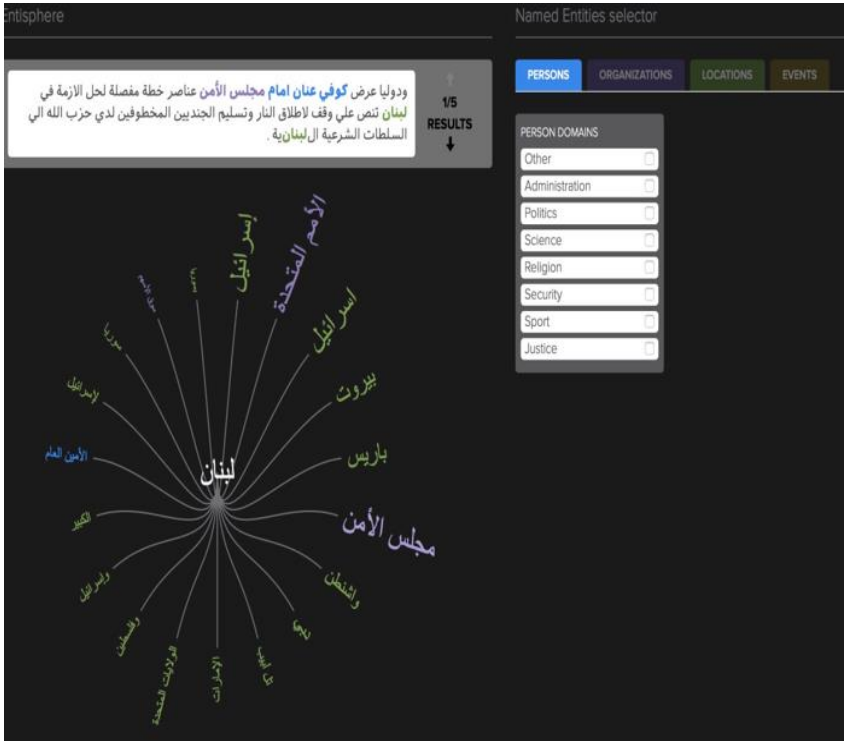


Figure 1. Résultat sur la requête لبنان (Liban)

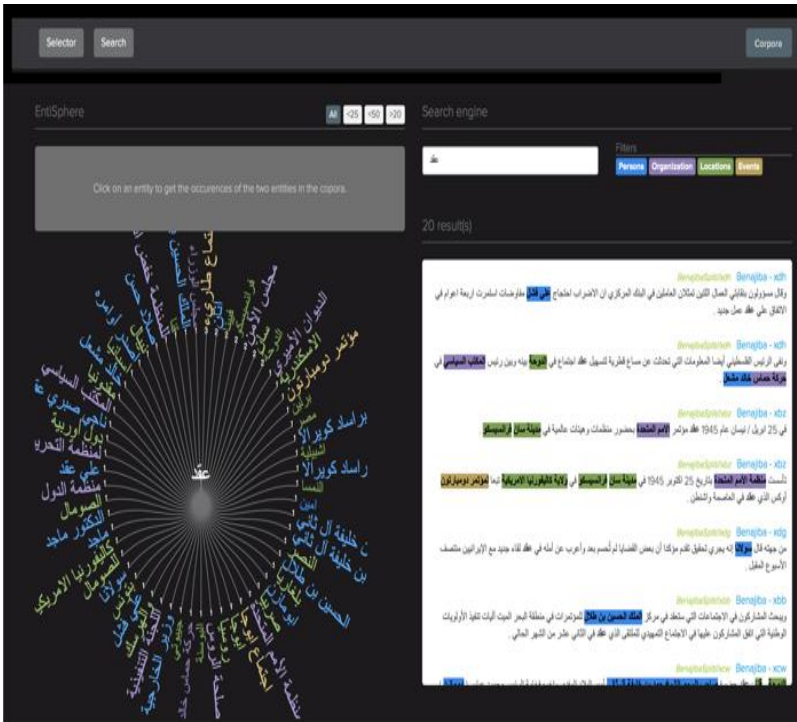


Figure 2. Résultat sur la requête عقد (contrat)

Ici, dans ce système de recherche d'informations, dont l'indexation est basée sur les résultats (sortie xml) de notre système de reconnaissance et de classification des EN en arabe, le filtrage peut se faire par classe (ex. PERSONNE), domaine (ex. Science) et nom propre (ex. الحسين بن طلال Al-Houssine Ibn Talal).

Comme on peut voir ici, la requête est entourée par des EN (attribut « nom propre » de différentes classes et attribut « fonction » de la classe «PERSONNE»). En cliquant sur une EN, le texte (à droite) affiche et marque les relations qu'a cette dernière avec les EN dans le corpus.

B) Extraction d'information :

Voici un exemple d'interface d'extraction d'informations :

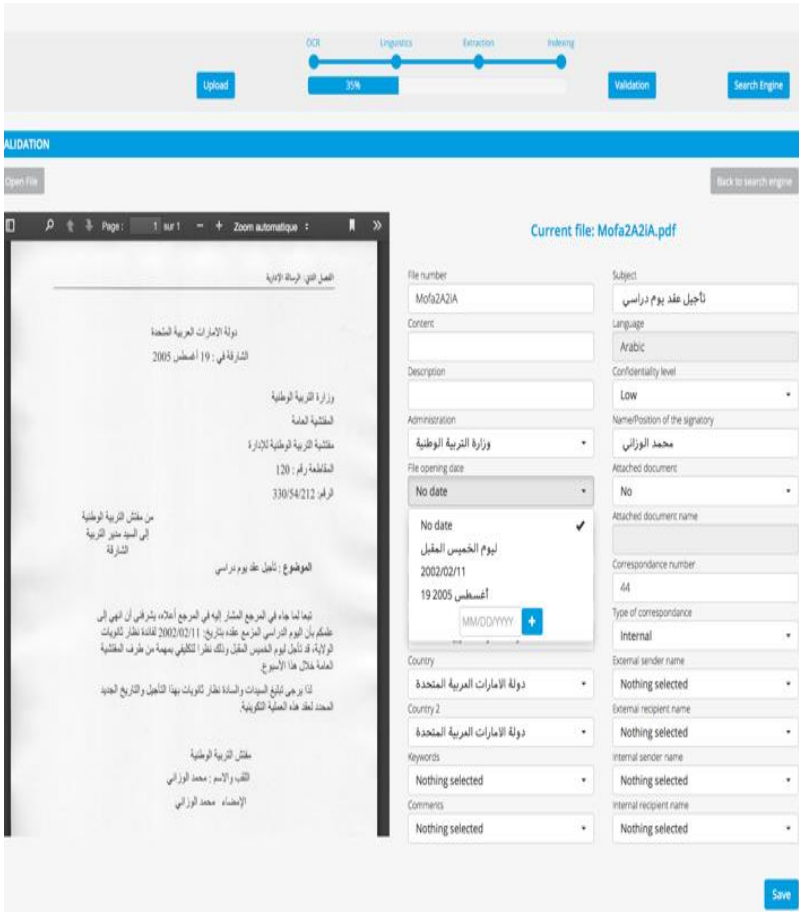


Figure 3. Interface d'extraction d'information à partir d'un courrier

Dans ce système d'extraction d'information, les champs sont remplis par les attributs « nom d'organisation », « date », « nom de pays » et « nom de personne ». Cette opération s'appuie sur la reconnaissance et la classification dans la sortie de notre système.

9 Conclusion :

Les EN sont des unités linguistiques à forte valeur informative. Leur reconnaissance contribue à une représentation sémantique du contenu textuel et constituent des éléments d'indexation importants et pertinents pour un moteur de recherche et aussi pour les systèmes d'extraction d'informations. Premièrement, il convient de reconnaître les concepts pour mieux les cibler. Deuxièmement, il est indispensable d'explorer leur représentation linguistique par des entités syntactico-sémantiques auxquelles sont associées des annotations sémantiques.

La détermination des attributs sémantiques des classes des EN sont essentiel dans la construction de leurs structures syntactico-sémantiques. Chaque attribut est représenté par un type bien déterminé d'entité linguistique : morphosyntaxique, syntaxique et sémantique. Il est donc primordial de les reconnaître, les extraire et les classer pour pouvoir ensuite construire, dans notre système à base de règles, les patrons syntactico-sémantiques de reconnaissance et de classification fine des entités nommées. Ces attributs nous sont également utiles dans la reconnaissance, dans la recherche d'informations et l'extraction d'informations, et des relations entre entités en exploitant la classification que nous leur avons attribuées. Pour un moteur de recherche le filtrage de l'information peut se faire par les différents attributs ; et pour l'extraction d'information, les champs recherchés peuvent être remplis par les entités correspondant à ces attributs.

10 Références :

- Abney, S. (1995). Chunks and dependencies: bringing processing evidence to bear on syntax". In *Computational Linguistics and the Foundations of Linguistic Theory, CSLI*.
- Aussenac-Gilles, N., Biebow, B., Szulman, S. (2000). "Revisiting ontology design: a methodology based on corpus analysis". In *12th International Conference, EKAW*, 27-66.
- Bourigault, D., Aussenac-Gilles, N. (2003). Construction d'ontologies à partir de textes. In *Actes de la 10ème conférence annuelle sur le Traitement Automatique des Langues*, 27-50.
- Charton, E. Gagnon, M. et Ozell, B. (2010). Extension d'un système d'étiquetage d'EN en étiqueteur sémantique. In *proceedings TALN 2010*, Montréal.
- Chomsky, N. (1967). The formal nature of language. In *E. H. Lenneberg, Biological Foundation of language*, Willey, New York.
- Chomsky, N. (1957). *Syntactic structures*. Mouton, La Haye.
- Cohen, D. (1961/70). Essai d'une analyse automatique de l'arabe, 1961 (T.A. informations). In *D. Cohen, Études de linguistique sémitique et arabe*, Mouton Paris, 49-78
- Cucchiarelli, A., Velardi P. (2001). Unsupervised named entity recognition using syntactic and semantic contextual evidence. In *Computational Linguistics*, 27(1), MIT Press, Cambridge, 123-131.
- Cusin-Berche, F. (2003). *Les mots et leurs contextes*. Paris : Presses Sorbonne nouvelle.
- Daille, B., Fourour N., et Morin E. (2000). « Catégorisation des noms propres : une étude en corpus ». In *Cahiers de Grammaire*, Vol 25, pp. 115-129.
- Depecker, L. (2002). *Entre signe et concept : éléments de terminologie générale*. Presses Sorbonne nouvelle, Paris.
- Desclés, J.-P., Abaab, H., Dichy J., Kouloughli D.E., Ziadah M.S. (1983). *Conception d'un synthétiseur et d'un analyseur morphologiques de l'arabe, en vue d'une utilisation en enseignement assisté par ordinateur*. Rapport rédigé sous la direction de J.-P. Desclés, à

la demande du Ministère français des Affaires étrangères (sous-direction de la Politique linguistique).

Dichy, J. (1984). Vers un modèle d'analyse automatique du mot graphique non-vocalisé en arabe. In *Dichy et Hassoun, eds.*, 92-158.

Dichy, J. (1987). The SAMIA Research Program, Year Four, Progress and Prospects. *Revue Processing Arabic Report n°2* (1987), T.C.M.O, Université catholique de Nimègue, 1-26.

Dichy, J. et Hassoun M.O. éd. (1989). *Simulation de modèles linguistiques et enseignement assisté par ordinateur de l'arabe* - Travaux SAMIA I, Conseil International de la Langue Française, Paris.

Dichy, J. (1990). *L'écriture dans la représentation de la langue : la lettre et le mot en arabe*. Thèse d'État (en linguistique), Université Lumière-Lyon2.

Dichy, J. (1997). Pour une lexicomatique de l'arabe : l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot. In *Meta, journal des traducteurs*, Montréal, Vol. 2, 223-474.

Dichy, J., Braham, A., Ghazali, S., Hassoun, M. (2002). La base de connaissances linguistiques DIINAR.1 (DIctionnaire INformatisé de l'Arabe, version 1). In *Proceedings of the International Symposium on The Processing of Arabic*, Université de Manouba, Tunis.

Dichy, J., Hassoun, M. (2005). The DIINAR.1-« معالي » Arabic lexical resource, an outline of contents and methodology. In *The ELRA Newsletter*, Vol. 10, num. 2, 5-10.

Ditters, E. (1992). *A formal approach to arabic syntax: the noun phrase and the verb phrase*. Thèse de doctorat, University Catholique de Nijmegen.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., Weischedel, R. M. (2004). The Automatic Content Extraction (ACE) program-tasks, data, and evaluation. In *LREC*.

Ehrmann, M., Jacquet, G. (2006). Vers une double annotation des EN. In *Traitement Automatique du Langues*, vol. 47, 63-88.

El Maarouf, I., Villaneau, J., Rosset, S. (2011). Extraction de patrons sémantiques appliquée à la classification d'EN. In *TALN*, Montpellier.

Farber, B., Freitag, D., Habash, N., et Rambow, O. (2008). Improving NER in Arabic Using a Morphological Tagger. In *Proceedings of LREC*, Marrakech.

Fleischman, M. (2001). Automated sub-categorization of named entities. In *ACL*, 25-30.

Fleischman, M., Hovy, E (2002). Fine grained classification of named entities. In *Proceedings of the 19th international conference on Computational linguistics*, Morristown, NJ, USA, 1-7.

Hassoun, M. O. (1987). *Conception d'un dictionnaire pour le traitement automatique de l'arabe dans différents contextes d'application*. Thèse d'état (en Sciences de l'information), Université Lyon 1.

- Hagène, C., Roux, C. (2003). Entre syntaxe et sémantique : normalisation de la sortie de l'analyse en vue de l'amélioration de l'extraction d'information à partir de texte. In *Actes de TALN 2003*, Batz-sur-Mer, 11-14.
- Jackendoff, R. (1977). *X-Bar syntax: a study of phrase structure*. The MIT Press, Cambridge (Mass.)
- Kaplan, R., Bresnan, J. (1982). Lexical-functional grammar: a formal system for grammatical representation. In *The Mental Representation of Grammatical Relations*, MIT Press, Cambridge, Massachusetts, États-Unis, pp. 173-281.
- Kleiber, G. (1981). *Problèmes de référence : descriptions définies et noms propres*. Klincksieck, Paris .
- Kleiber, G. (1990). *La sémantique du prototype : catégories et sens lexical*. Presses universitaires de France, Paris.
- Knopp, J. (2010). *Classification of named entities in a large multilingual resource using the Wikipedia category system*. Thèse de doctorat, Université de Heidelberg.
- Le Pesant, D., Mathieu-Colas, M. (1998). Introduction aux classes d'objets. In *Langages*, 32e année, num. 131, Les classes d'objets, 6-33.
- Lecolle, M., Paveau, M. A., Reboul-Touré, S. (2009). Les sens des noms propres en discours. In *Les Carnets du Cediscor*, 11. Publication du Centre de recherches sur la didacticité des discours ordinaires, 9-20.
- Lyons, J. (1977). *Semantics*. Presses universitaires de Cambridge, Cambridge.
- Martineau, C., Tolone, E., et Voyatzi, S. (2007). Les EN : usage et degrés de précision et de désambiguïsation. In *26ème Colloque international sur le Lexique et la Grammaire*, 105-112.
- Maurel D., Tran M., Vitas D., Grass T., Savary A. (2006). Prolex : Implantation d'une ontologie multilingue des noms propres. *Rapport interne du Laboratoire d'Informatique de l'Université François-Rabelais de Tours*, num. 286, 47 pages.
- Moortgat, M. (1987). *Compositionality and the syntax of words*. Groningen-Amsterdam *Studies in Semantics*, (7), 41-62.
- Ouersighni, R. (2002). *La conception et la réalisation d'un système d'analyse morpho-syntaxique robuste pour l'arabe : utilisation pour la détection et le diagnostic des fautes d'accord*. Thèse en Science de l'Information et de la Communication, université lumière-Lyon2.
- Paik, W., Liddy, E. D., Yu E., Mckenna, M. (1996). Categorizing and standardizing proper nouns for efficient information retrieval. In *Corpus Processing for Lexical Acquisition*, Massachusetts Institute of Technology, 61-73.
- Poibeau, T. (2003). *Extraction automatique d'information : Du texte brut au web sémantique*. Hermès, Paris, 250 pages.

- Poibeau, T. (2011). *Traitement automatique du contenu textuel*. Hermes science publications, Lavoisier, Paris.
- Polguère, A. (2003). *Lexicologie et sémantique lexicale ; Notions fondamentales*. Les Presses de l'Université de Montréal, Montréal.
- Roman, A. (1983). Sur la constitution des unités de la langue arabe. In *Travaux 1 du Cercle linguistique d'Aix-en-Provence, Les Parties du discours*, Aix-en-Provence, 113-144.
- Roman, A. (1999). *La création lexicale en arabe*. Presses universitaires de Lyon, Lyon.
- Tesnière, L. (1959). *Éléments de syntaxe structurale*. Klincksieck, Paris.
- Traboulsi, H. (2009). Arabic named entity extraction: a local grammar-based approach. In *Proceedings of the International Multiconference Computer Science and Information Technology*, 139-143.
- Trudel, E. (2009). Champ sémantique, champ lexical ou classe sémantique ? Traitement conceptuel en lexicologie et en sémantique. In *Texto*, disponible à : <http://www.revue-texto.net/index.php?id=2277>
- Zaghouani, W. (2009). *Le repérage automatique des EN dans la langue arabe : vers la création d'un système à base de règles*. Mémoire de Maîtrise, Université de Montréal.