RECONNAISSANCE DES ENTITES NOMMEES EN ARABE

Sylvie GUILLEMIN-LANNE (*), Fathi DEBILI (**), Zied Ben TAHAR (**), Chafik GACI (*)
sylvie.guillemin-lanne@temis.com, fathi.debili@wanadoo.fr, bentaharzied@gmail.com, chafik.gaci@temis.com

(*) TEMIS, Tour Gamma B, 193-197 rue de Bercy, 75012 PARIS, France
(**) LLACAN, INALCO, CNRS, 7, rue Guy Môquet, 94801 Villejuif, France

Mots clefs :
Veille scientifique et technologique, Fouille de données textuelles, ingénierie des connaissances, extraction d'information, entités nommées, règle d'extraction, patron d'extraction, intelligence économique**,** ingénierie des connaissances, modélisation des connaissances
Keywords:
Scientific and technical observation, text mining, knowledge engineering, knowledge extraction, information extraction, named entities, extraction rules, extraction pattern., competitive intelligence, knowledge engineering, knowledge modeling
Palabras clave :
Escudriñar científico y tecnológico text mining, ingeniería del conocimiento, extraccion del conocimiento, extraccion de la informacion, reglas de extraccion, ingeniería del conocimiento, formalización del conocimiento

Résumé

Cet article décrit une application à la langue arabe des technologies de text mining développées au sein de la société TEMIS. L'objectif est de développer une chaine de reconnaissance des entités nommées en arabe. Après une étude synthétique des spécificités de la langue et des entités nommées en arabe, l'article met l'accent sur les capacités du taggueur et les choix méthodologiques à résoudre efficacement les ambiguïtés dues à la langue. Une partie est réservée à l'extraction des entités nommées proprement dites, en utilisant les technologies existant pour d'autres langues et/ou en les adaptant.

1 Introduction

Contraction de Text Mining Solutions, TEMIS est un éditeur de logiciel qui conçoit et propose des applications multilingues dédiées à l'analyse textuelle, et liées à l'intelligence économique, à la gestion de la relation clients, à la gestion de la connaissance et des savoir-faire et à la gestion des ressources humaines. La demande est forte pour intégrer de nouvelles langues stratégiques au sein de nos applications telles que l'arabe, par exemple.

Notre nouveau challenge a pour objectif de reconnaître des entités nommées en arabe. Pour cela nous devons éprouver la capacité des outils Temis et du formalisme des Skill Cartridge™ à traiter la langue arabe. Après une étude des spécificités de la langue arabe et des entités nommées en arabe, nous listerons les caractéristiques techniques du taggueur de l'arabe que nous avons choisi ainsi que la méthodologie utilisée pour gérer les nombreuses ambiguïtés.

Enfin nous aborderons notre dernière partie sur l'extraction des entités nommées en arabe. Cette partie fait l'objet des perspectives que nous souhaitons mettre en œuvre. Au vu des résultats d'annotation des entités nommées nous déciderons des techniques d'extraction d'information que nous souhaitons reprendre ou adapter de l'existant, ainsi que les nouvelles stratégies à mettre en œuvre. Nous appuierons notre propos par une validation qualitative et quantitative des entités extraites sur différents types de corpus. La qualité des résultats attendus nous permettra d'asseoir nos choix technologiques et méthodologiques.

2 Les entités nommées en arabe

2.1 Les caractéristiques de la langue arabe

« The general rule of thumb for Arabic is that everything is at least five times more complicated than for any European language » ***Kenneth R. Beesley.***
La principale caractéristique de la langue arabe vient du fait qu'il s'agit d'une langue non voyellée ; hors contexte, il est difficile de distinguer le sens et la fonction des mots. Cette caractéristique introduit, de fait, une forte ambiguïté avec laquelle il va falloir « jouer » dans le cadre du traitement automatique de la langue.

2.1.1 L'écriture de droite à gauche

Traiter automatiquement la langue arabe suppose d'utiliser un éditeur de texte capable de soutenir l'écriture de droite à gauche et surtout, de combiner les différents sens d'écriture (par exemple, du français ou de l'anglais avec de l'arabe). Il est en effet courant de trouver dans les textes arabes, des noms écrits en langue latine avec leur acronyme et, à côté, de trouver la traduction du nom de la société en arabe.

2.1.2 La voyellation

Contrairement à la langue française, les voyelles dans la langue arabe ne sont pas des lettres, mais des signes qui s'écrivent au dessus ou au dessous des lettres (consonnes) et qui remplissent la fonction de voyelle.

A titre d'exemple, prenons le mot كتب / ktb et comptabilisons ses diverses voyellations [Debili, 2002] :

« كَتَبَ / kataba » (Il a écrit)

« كُتِبَ / kutiba » (Il a été écrit)
« كُتُب / kutub » (des livres)
« كَتْب / katb » (un écrit)
« كَتَّبَ / kattaba » (Il a fait écrire)
« كُتَّبَ / kuttiba » (faire écrire - forme factitive)
« كَتِّبْ / kattib » (fais écrire)

Les mots arabes acceptant plusieurs voyellations, la tâche de désambiguïsation est primordiale. Cette problématique est double [Debili, 2002] :

1. Comment étiqueter un texte non voyellé alors que la détermination des voyelles semble devoir précéder celle des étiquettes grammaticales, puisque ce sont elles qui aident à déterminer ces dernières ?
   C'est par exemple la connaissance de la ré-accentuation potentielle *élevé* de *eleve* qui permet d'attacher l'étiquette potentielle « participe passé » à *eleve*.
2. Comment restituer les voyellations respectives de chacun des mots d'un texte, alors que, inversement, la détermination de celles-ci semble dépendre de la détermination des étiquettes grammaticales ?
   C'est par exemple le choix de l'étiquette « participe passé » qui force la sélection de la forme accentuée *élevé* et non de *élève* toutes deux attachées à *eleve*.

2.1.3 L'agglutination

Contrairement aux langues latines, l'arabe est une langue agglutinante. Les articles, les prépositions et les pronoms collent aux adjectifs, noms, verbes; ce qui nécessite de procéder au découpage des mots avant la tâche de lemmatisation. La plupart des mots arabes sont composés par l'agglutination d'éléments lexicaux élémentaires. Par exemple, la détermination peut s'exprimer par :

- l'agglutination de l'article ال AL avant le mot.

  Le livre : الكتاب

- l'agglutination d'un clitique à la fin du mot :

  Son livre : كتابه

La forme agglutinée correspond à une suite de formes « collées ».

Les ambiguïtés segmentales

L'agglutination suppose une segmentation des unités morphologiques, ce qui va engendrer de multiples ambiguïtés, sachant qu'un texte non voyellé peut avoir un nombre important de segmentations possibles, celles-ci n'ayant pas toutes le même sens.
Dans sa forme non voyellée المهم ('lmhm), le même mot, accepte au moins les trois segmentations suivantes :

- أ + لمّ + هم ('+*lmm*+*hm* les a-t-il ramassés)
- ألم + هم ('*lm*+*hm* leur douleur ou *'llm*+*hm* il les a fait souffrir)

- الـ + مهم (*’l+mhm* l’important)

Il est presque toujours possible de trouver plusieurs segmentations valides pour un seul mot non voyellé. Pour lever l’ambiguïté, il faut avoir des connaissances supplémentaires, qu’elles soient sémantiques ou/et syntaxiques.

Il n’en est pas de même pour un texte voyellé.
Le mot أَلَمُهُمْ (’alamuhum, leur douleur) dans sa forme voyellée n’accepte qu’une seule segmentation : أَلَمُ + هُمْ (’alamu+hum).

Pour distinguer les unités lexicales dans les unités morphologiques, il faudrait procéder à une analyse morphologique permettant de découper le mot en ses unités lexicales élémentaires. A chaque unité morphologique, il faudrait attribuer une ou plusieurs unités lexicales, permettant de distinguer l’article défini du nom, comme dans : الـ + مهم (’l+mhm l’important) et distinguer le nom du clitique comme dans : ألم + هم (’lm+hm leur douleur ou ’llm+hm il les a fait souffrir) etc…
Les unités à extraire doivent être connues et répertoriées dans des lexiques. Pour ce faire, il est nécessaire d’avoir une segmentation valide.

<u>Exemple de segmentation valide :</u>

- Puis il est venu : فجاء
  ف: FA conjonction de coordination (puis) qui a une fonction antéfixe.
- جاء: verbe جاء (il est venu)

<u>Exemple de segmentation invalide :</u>
Il a ouvert : فتح

En effet, si l’on segmente le mot comme : ف FA et تح TAHA, la segmentation sera fausse ; Séparé ainsi le mot n’a aucun sens car, dans ce cas, le mot fataha n’est pas formé par agglutination, une forme n’a pas toujours la même fonction, elle peut être un mot ou une partie de mot (ici une consonne faisant partie du mot).

Le tableau suivant donne pour un texte arabe d’environ 25000 mots les proportions d’unités morphologiques (UM) acceptant respectivement une seule ou plusieurs segmentations.

	Nb UM	**UM Non Ambiguës**	**UM Ambiguës**	**Seg/UM**	**Nb max de seg.**
Voyellé	25 410	96,61 %	3,39 %	1,03	4
Non voyellé	25 410	78,00 %	22,00 %	1,30	6

Tableau 1: unités morphologiques donnant lieu à des segmentations en proclitique + forme simple + enclitique ambiguës [Debili, 2002]

Sous l'angle de l'agglutination, on remarque donc que la segmentation d'un texte non voyellé est bien plus ambiguë que celle de son correspondant voyellé :

- Le nombre d'unités admettant plus d'une segmentation est d'abord plus important : 22% contre 3,39 %.
- De plus, le nombre moyen de segmentations par unité est plus grand pour le non voyellé que pour le voyellé : 1,3 segmentation en moyenne contre 1,03 pour le voyellé.

Le tableau 1 indique en outre que le nombre maximal de segmentations observées est de 4 pour le voyellé et de 6 pour le non voyellé. [Debili, 2002]

L'analyse morphologique devra donc séparer et identifier des morphèmes semblables aux mots préfixés comme les conjonctions wa- و et fa- ف, des prépositions préfixées comme bi- ب et li- ل, l'article défini ال , des suffixes de pronom possessif.

2.1.4 Les variantes graphiques

Dans l'alphabet arabe, chaque lettre possède quatre allographes1, à l'exception d'un petit nombre de lettres dont le tracé reste invariable. Chaque variante s'utilise dans un contexte précis dépendant de sa place dans le mot :

- en position indépendante, lorsque la lettre est seule dans le mot. زرع → il a semé.
- en position initiale d'un mot. عمل → il a travaillé.
- en position médiane. يعمل → il travaille.
- en position finale. أصبع → un doigt.

Remarque : Pour certaines lettres le rattachement par la gauche à d'autres lettres est impossible, ce sont les lettres suivantes :

ا : *A*
د : *D*
ذ : *DH*
ر : *R*
ز : *Z*
و : *W*

2.1.5 L'absence de majuscule

Contrairement à d'autres langues comme le français ou l'anglais, la langue arabe ne dispose pas de la notion de majuscule. Sachant que la majuscule est un moyen très efficace dans le processus de reconnaissance des noms propres, son absence dans la langue arabe nous contraint à trouver d'autres solutions ou, à la limite, d'utiliser d'autres moyens classiques comme les lexiques, les mots amorces et les règles grammaticales.

2.1.6 Bilan de l'étude préliminaire

Cette étude préliminaire sur les spécificités de la langue a été une étape nécessaire afin de définir les prérequis d'un système de traitement automatique de l'arabe. A savoir :

[1] Allographe, ou façon de s'écrire.

- L'écriture droite-gauche : l'éditeur de texte devra être capable de soutenir l'écriture de droite à gauche et combiner les différentes sens d'écriture (exemple : combiner du français ou de l'anglais avec de l'arabe) ;
- L'absence des voyelles, qui engendre une ambiguïté importante des textes non voyellés ;
- L'agglutination, qui augmente la difficulté de la lemmatisation par l'obligation d'un découpage préalable des mots.
- L'absence de majuscule qui prive le taggueur d'un moyen efficace de reconnaissance de suites de noms propres.

2.2 Le choix d'un taggueur

Pour procéder à l'analyse morpho-syntaxique de l'arabe, nous avons cherché un taggueur qui sache traiter les difficultés liées à la langue arabe exposées dans le paragraphe précédent, et ce de façon performante en terme de qualité.. Dans un objectif d'industrialisation, nous avons ajouté les critères suivants :

- Couverture : L'analyseur doit pouvoir analyser la langue arabe écrite, quel que soit le contexte
- Robustesse : Quels que soient la volumétrie et le format des textes sources, l'analyseur doit pouvoir fournir des résultats reproductibles
- Performances : L'analyseur doit être capable d'analyser des documents dans des temps acceptables
- Accessibilité : L'analyseur considéré doit être ouvert et permettre l'accès aux ressources linguistiques.
- Maintenabilité : L'analyseur considéré doit être capable d'évoluer en fonction des retours de tests. Les équipes techniques et linguistiques doivent être présentes et assurer son l'évolutivité.

Avant de procéder au choix de G-Lexar, nous avons évalué le taggueur sur un corpus du Monde Diplomatique acquis auprès de ELDA[2] et préalablement annoté à la main par l'équipe qui a développé le taggueur. Au cours de cette tâche, un autre taggueur a également été évalué, dont les résultats ne seront pas publiés.

2.2.1 Le taggueur G-Lexar

G-Lexar est un analyseur morpho-grammatical de l'arabe issu des travaux menés au CNRS sur le traitement automatique de l'arabe depuis plus de 25 ans.

2.2.1.1 Les étapes d'analyse

L'analyse morpho-grammaticale est modulaire et séquentielle. Elle s'effectue selon les étapes suivantes :

1. Segmentation du texte en unité morphologique
2. Analyse morphologique

Pour segmenter les unités morphologiques, l'analyse utilise différents dictionnaires brièvement décrits ci-dessous ainsi que plusieurs grammaires qui portent sur les compatibilités proclitiques, radical, enclitiques, d'une part, et sur des règles de réécritures permettant de prendre en charges diverses variations morphologiques (graphies multiples de la hamza « ا أ إ », harmonie vocalique liée aux pronoms clitiques, etc.), d'autre part.
Le but est la production d'analyses morphologiques – celles-ci se présentent sous une forme arborescente (découpages potentiels, voyellations potentielles, lemmes potentiels, étiquettes potentielles), tout en éliminant des segmentations proclitique+forme simple+enclitique illicites.

[2] ELDA: Corpus de textes du journal "Le Monde Diplomatique" en arabe- Années 2001 à 2004 (ELRA-W 003604).

2.2.1.2 Dictionnaires mis en œuvre

Les dictionnaires se décomposent comme suit :

- Proclitiques et enclitiques (66 entrées),
- Formes simples :
  - 500 000 entrées environ non voyellées,
  - 1 500 000 entrées simples voyellées
  - 82 000 lemmes environ.
- Les lemmes sont voyellés et accompagnés de leurs schémas casuels et étiquettes grammaticales hors contexte.

2.2.1.3 Les jeux d'étiquettes

Plusieurs jeux d'étiquettes grammaticales sont mis en œuvre :

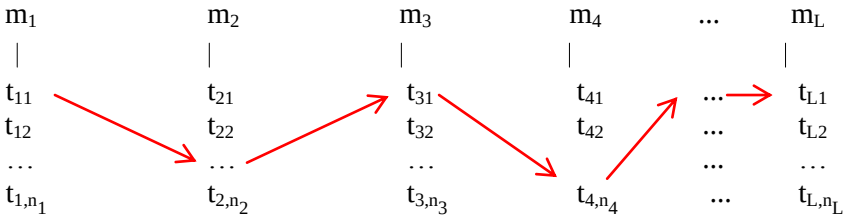
- 250 étiquettes simples,
- 1 730 étiquettes simples et composées

La lecture des étiquettes est explicite. Nous les dénommons hyper catégories grammaticales car elles se rapportent aux formes simples ou agglutinées de l'arabe. Une description détaillée est donnée dans (Debili F., Achour H., Souissi E. 2002).

2.2.1.4 Méthodologies d'étiquetage

L'étiquetage dans G-Lexar est fondé sur l'utilisation de règles qui portent sur les successions permises non pas de deux ou trois étiquettes grammaticales, ainsi qu'il est traditionnellement fait, mais sur les successions de listes d'étiquettes, successions qui vont de deux à 11 listes [Debili et al. 2005]. Rappelons que l'objectif de l'étiquetage grammatical consiste à attribuer à chacun des mots du texte la catégorie qui est la sienne dans le contexte où ce mot apparaît.

La figure suivante rappelle ce qui se passe sous l'angle combinatoire :



où mi sont les mots de la phrase, et tij les étiquettes grammaticales potentielles respectivement associées aux mi.
Ces règles se présentent formellement comme suit. L'exemple suivant illustre le cas des règles de succession de deux listes :
$\{t_{11}, t_{12}, t_{1n_1}\}$ $\{t_{21}, t_{22}, t_{2n_2}\}$ → $(t_{1i_1}, t_{2j_1}) \mid (t_{1i_2}, t_{2j_2}) \mid (t_{1i_3}, t_{2j_3}) \mid$ etc.
Que l'on peut lire ainsi :
La liste d'étiquettes $\{t_{11}, t_{12}, t_{1n_1}\}$ suivie de la liste $\{t_{21}, t_{22}, t_{2n_2}\}$ peuvent être respectivement résolues en t_{1i_1}, pour la liste 1, et t_{2j_1}, pour la liste 2 ou, de façon plus concise, par (t_{1i_2}, t_{2j_2}), ou par (t_{1i_3}, t_{2j_3}), etc.

- Ces règles permettent d'éviter l'explosion combinatoire liées à l'utilisation de règles de succession de simples étiquettes. Rappelons en effet que les successions de listes d'étiquettes, outre qu'elles sont plus discriminantes, renferment d'emblée les successions licites d'étiquettes, rendant ainsi inutile le développement combinatoire des différents trajets syntaxiques qu'engendre l'utilisation des règles de successions d'étiquettes simples.

Ces règles ont été proposées il y a plus de trente ans par A. Andreewsky et C. Fluhr. Le peu de succès qu'elles ont eu est lié au fait qu'elles réclament des corpus d'apprentissage bien plus importants, ce qui n'est plus un frein aujourd'hui étant donné la puissance des machines. En effet, la taille des corpus d'apprentissage permettent d'envisager des règles qui ne portent plus seulement sur des successions de deux ou trois listes, et de les chaîner lors de leur utilisation.

2.2.2 Evaluation du taggueur G-Lexar

2.2.2.1 Le corpus d'évaluation

Le corpus d'évaluation est un corpus annoté de 15 296 mots (qui correspondent à 9 textes du Monde Diplomatique) voyellés, lemmatisés et étiquetés manuellement.

G-Lexar stocke le résultat dans un fichier XML dont la structure est la suivante :

```
<mot NumParagraphe="1" NumPhraseInParagraphe="1" NumPhrase="1" rang="2" position="4" tailleMot="6">
  <UM>الهاتف</UM>
  <Voyellation>الْهَاتِف</Voyellation>
  <Lemme>هَاتِف</Lemme>
  <Categorie>Substantif</Categorie>
</mot>
```

Figure 1 : Exemple de la structure XML du corpus d'évaluation

Chaque nœud contient les « attributs » positionnels du mot dans le texte :

- NumParagraphe: désigne le numéro du paragraphe dans le texte dont le mot est issu.
- NumPhraseInParagraphe: numéro de la phrase dans le paragraphe dont le mot est issu.
- NumPhrase: désigne le numéro de la phrase dans le texte.
- rang: désigne le rang du mot dans le texte.
- Position: désigne le rang du premier caractère du mot dans le texte.
- TailleMot : désigne la taille du mot.
- Les informations concernant l'annotation du mot sont les sous nœuds :
- UM: pour Unité Morphologique; mot non voyellé.
- Voyellation: le mot voyellé.
- Lemme: le lemme du mot.

- Catégorie: la catégorie grammaticale du mot.

2.2.2.2 Le choix de l'étiquette grammaticale

Pour optimiser les résultats, G-Lexar propose pour une unité morphologique plusieurs choix d'étiquettes grammaticales.
Cette démarche, de nature linguistique, se base sur la connaissance de l'étiquette du ou des mots précédents, sur les règles grammaticales et sur l'ordre des mots dans la phrase pour réduire l'étendue du choix d'une étiquette ou déduire une étiquette grammaticale. Il s'agit tout simplement de l'application du principe des règles de succession sur lesquelles est basé l'algorithme de G-Lexar, qui sont générées automatiquement par apprentissage sur des corpus de textes.

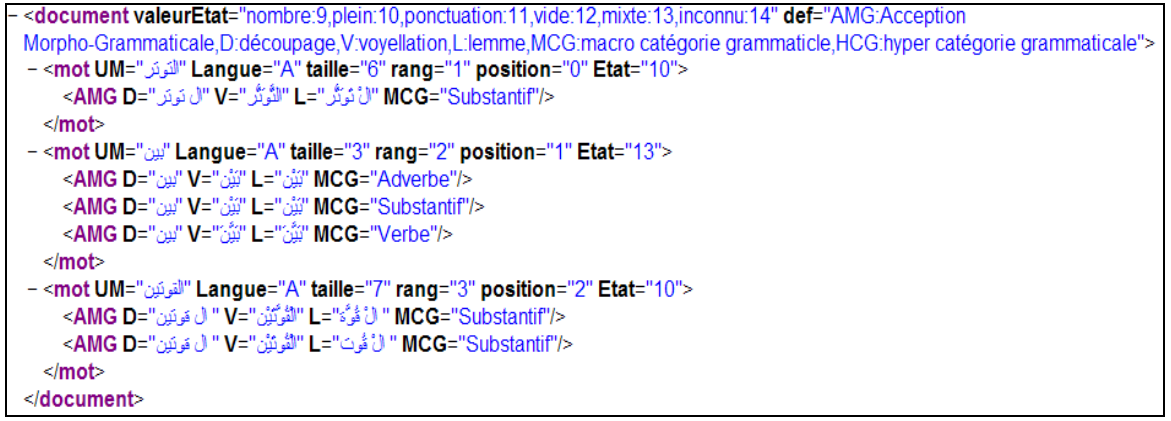


Figure 2 : Résultat d'une analyse grammaticale effectuée par G-Lexar

2.2.2.3 Résultats d'évaluation de G-Lexar

Description du corpus de test

Nom du corpus	Monde Diplomatique
Nombre de mots du corpus	15 297 mots
Nombre de mots qui n'ont pas été reconnus par l'analyseur morphologique	611 mots
Taille Corpus d'apprentissage	236 000 mots

Performances Etiquetage : les comptages se rapportent d'une part aux catégories fines (663 catégories effectivement attribuées), d'autre part à des macro-catégories (37 au total)

Proportion des mots correctement étiquetés en tenant compte des mots non reconnus	Etiquetage fin	67,14%
	Macro étiquettes	84,72%
Proportion des mots correctement étiquetés sans tenir compte des	Etiquetage fin	69,93%

mots non reconnus	Macro étiquettes	88,25%

Performances de la lemmatisation (après étiquetage grammatical)

	Etiquetage	Lemmatisation
Proportion des mots correctement lemmatisés en tenant compte des mots non reconnus	67,14%	78,07%
Proportion des mots correctement lemmatisés sans tenir compte des mots non reconnus	69,93%	81,77%

Ce tableau indique que 78% des mots sont correctement lemmatisés en se basant sur l'étiquette grammaticale et en faisant intervenir la fréquence relative f(lemme | mot non voyellé, étiquette). La lemmatisation qui ne fait pas intervenir l'étiquetage donne une proportion de lemmatisation correcte de 74,14%. Nous constatons que l'étiquetage grammatical préalable amène à une amélioration des résultats de lemmatisation d'environ 4%.

Performances de la voyellation (après étiquetage grammatical)

	Etiquetage	Voyellation
Proportion des mots correctement voyellés en tenant compte des mots non reconnus	67,14%	71,46%
Proportion des mots correctement voyellés sans tenir compte des mots non reconnus	69,93%	74,85%

Ce tableau indique que 71,46% des mots sont correctement voyellés en tenant compte de l'étiquette grammaticale et en faisant intervenir la fréquence relative f (voyellation | mot non voyellé, étiquette). La voyellation qui ne fait pas intervenir l'étiquetage donne une proportion de voyellations correctes de 69,76%. Nous constatons que l'étiquetage grammatical préalable amène à une amélioration des résultats liés à la voyellation d'environ 1,7%.

2.3 Expérimentations pour l'extraction des entités nommées

2.3.1 Etude sur les entités nommées à partir des corpus annotés

Une étude complémentaire a été menée sur les entités nommées. Le mode opératoire pour cette annotation semi-automatique s'est déroulé comme suit :

- Segmentation : délimitation manuelle des entités
- Typage manuel des entités nommées
- Extraction des successions d'étiquettes associées à ces syntagmes

- Parallèlement, extraction des successions de listes d'étiquettes associées à ces syntagmes.

La question sous-jacente est de savoir s'il existe une syntaxe des noms propres, si l'on peut déduire, par type d'entités nommées, une syntaxe minimale pour l'arabe.

Sur le plan méthodologique, nous voulons circonscrire les difficultés et moyennant un grand nombre d'entités nommées manuellement annotées et étiquetées, élaborer des règles morpho-syntaxique permettant de les reconnaître automatiquement.

Nous donnons ici un bref compte rendu chiffré de l'opération d'annotation manuelle des entités nommées arabes qui a été menée. Elle a porté sur un corpus d'environ 60 000 mots.

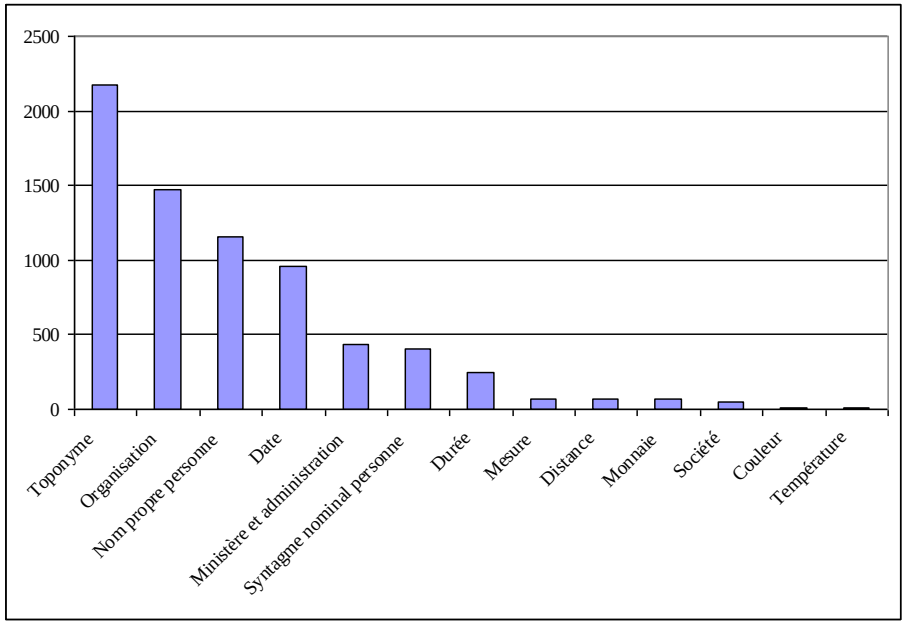


Figure 3*: histogramme des entités nommées.*

Interface d'annotation interactive des entités nommées

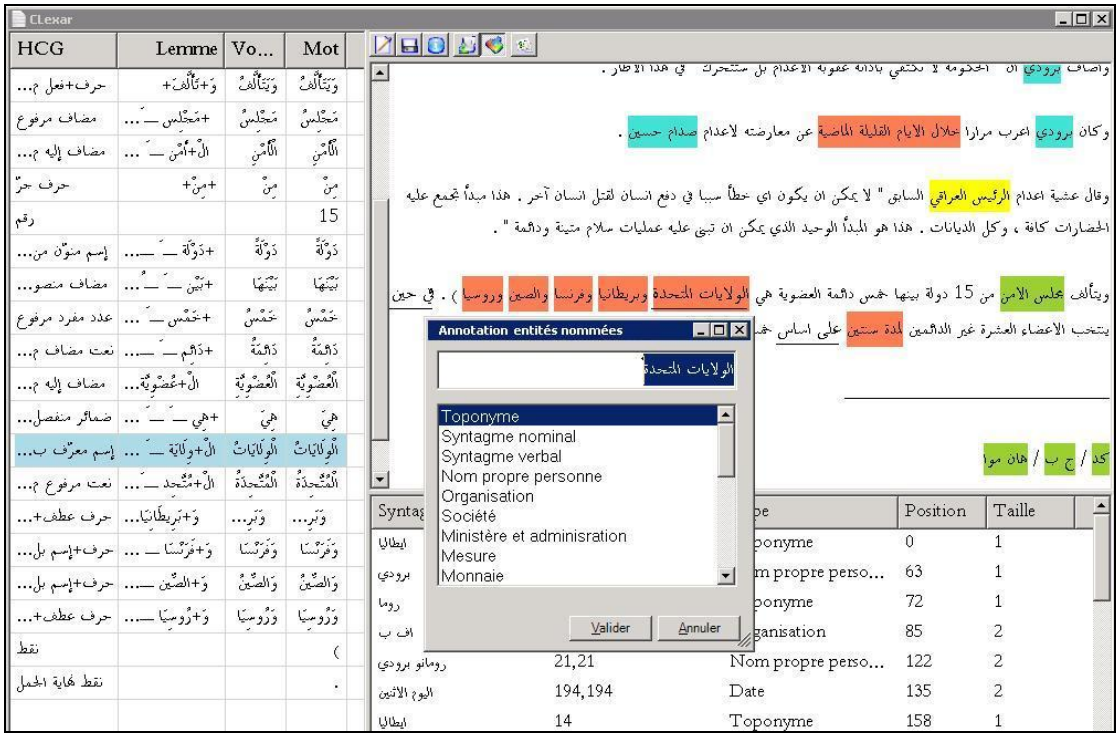


Figure 4: Outil d'annotation interactif.

L'annotation des entités nommées fait intervenir un filtre grammatical, permettant de suggérer la nature de l'entité nommée. Il simplifie et facilite l'annotation du type de l'entité manuellement délimitée. On trouvera :

- En haut, à droite : le texte après annotation
- En bas, droite : les entités nommées accompagnées d'un certain nombre d'informations (type, position, taille, hyper-catégories associées aux constituants de l'entité).

Entité	*Succession d'étiquettes*	*Type de l'entité*
يوم الخميس المقبل	167, 197,200	Date
يوم الجمعة	169, 197	Date
يوليانا فيداك	22,22	Nom propre personne

يوسف احمدي	21,21	Nom propre personne
يوروستات	22	Ministère et administration
يوروبا برس	21,22	Organisation
يورو	21	Monnaie

Tableau 2: Exemples d'entités nommées extraites à partir du corpus

2.3.2 Annotation des entités nommées sur une base lexicale

Parallèlement a été menée, sur la base des mêmes corpus, une étude sur l'annotation automatique des entités nommées à partir de lexiques uniquement. Le serveur d'extraction d'information Insight Discoverer™ Extractor, couplé à une Skill Cartridge™ construite à partir de lexiques arabes, a été lancé sur le même jeu de documents du Monde Diplomatique.
Un outil de validation des annotations, Skill Cartridge™ Monitor a été utilisé pour procéder à la validation des annotations.

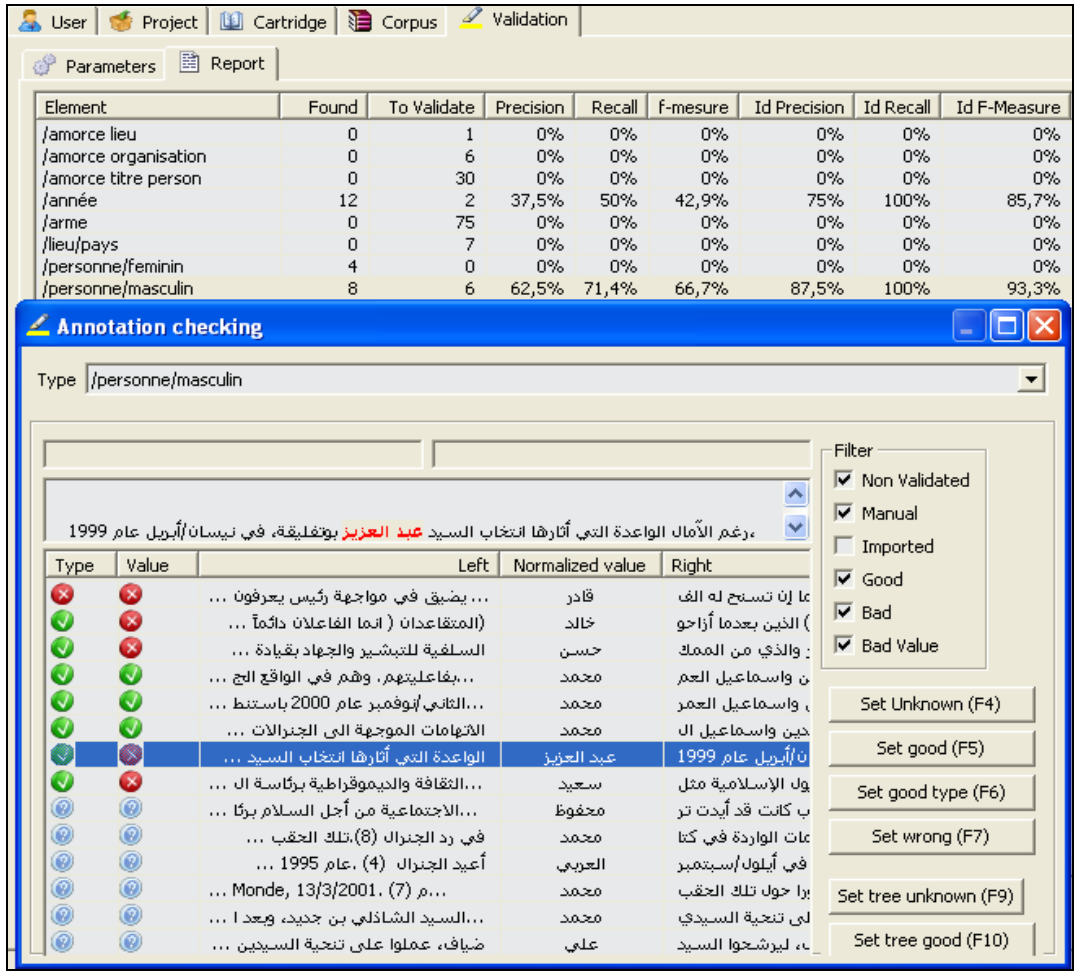


Figure 5: Validation des annotations avec Skill Cartridge™ Monitor.

Les résultats de l'annotation sont résumés dans le tableau ci-dessous :

Concepts	Total	Precision	Rappel	f-mesure	id-precision	id-rappel	id-f-mesure
/Amorce lieu	205	74,4%	83,3%	78,6%	89,3%	100,0%	94,3%

/Amorce organisation	345	79,3%	96,6%	87,1%	82,0%	100,0%	90,1%
/Amorce titre personne	804	81,9%	93,4%	87,3%	87,7%	100,0%	93,4%
/Annee	875	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
/Arme	2559	38,8%	85,3%	53,3%	45,5%	100,0%	62,5%
/Jour	6	33,3%	100,0%	50,0%	33,3%	100,0%	50,0%
/Lieu/Pays	853	99,4%	100,0%	99,7%	99,4%	100,0%	99,7%
/Mois/Mois Hegir	1	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
/Mois/Mois autre	34	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
/Mois/Mois latin	18	94,4%	100,0%	97,1%	94,4%	100,0%	97,1%
/Organisation	124	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
/Personne/Feminin	119	2,5%	100,0%	4,9%	2,5%	100,0%	4,9%
/Personne/Masculin	302	26,0%	55,7%	35,4%	46,7%	100,0%	63,7%
Global	**6245**	**63,9%**	**85,7%**	**68,7%**	**67,8%**	**92,3%**	**73,5%**

Tableau 3 : Résultats de la validation des entités nommées

De cette étude, il ressort que certaines entités comme les noms de pays, les mois, les organisations ont d'excellents taux de reconnaissance. Pour eux, la recherche lexicale s'avère très efficace. Ce n'est pas le cas des noms de personnes ou d'armes qui sont trop ambigus pour pouvoir être reconnus sur une base lexicale uniquement.

Ce constat pourra être nuancé, lorsque l'on ajoutera à notre test des patrons d'extraction construits à partir des mots amorces. Nous avons voulu, dans un 1er temps, tester la recherche lexicale par type d'entité, avant d'étendre notre test aux patrons d'extraction. Ceci étant, tous les noms de personne ne se construisent pas systématiquement avec un mot amorce.

3 Conclusion et Perspectives

L'objectif de notre projet est de reconnaître les entités nommées (personnes, lieux, organisations et sociétés, dates, emails,…) dans des textes arabes. Pour ce faire, nous couplons un taggueur de l'arabe au serveur d'extraction d'information de Temis. Nous avons choisi une approche qui s'appuie sur les résultats d'expérimentation pour avancer pas à pas dans la reconnaissance des entités nommées.

Cette approche est très utile compte tenu de l'augmentation croissante de l'information à analyser en langue arabe. Elle nous permet de tester nos performances à la fois sur le tagging et sur les résultats d'extraction d'information et de valider la faisabilité technique de l'analyse de la langue arabe avec les outils Temis, couplés à un taggueur performant de la langue arabe.. Alors que la demande est forte sur les outils capables d'analyser la langue arabe, nous serons fiers de présenter un retour d'expérience sur le domaine.

Bibliographie

[1] [Appelt *et al.* 1993] Appelt D., Hobbs J., Bear J., Israel D., Kameyama M. et Tyson M. « FASTUS : a finite-state processor for information extraction from real-world text ». I*n proceedings of the International Joint Conference on Artificial Intelligence* (IJCAI'93), Chambéry, 1993, pp. 1172-1178.

[2] [Aubry *et al*. 2002] Aubry Christophe, Grivel Luc, Guillemin-Lanne Sylvie, Lautier Christian « Aide à la construction de composants de connaissance pour l'extraction d'information : méthodologie et environnement» CIFT 2002 Colloque International sur la Fouille de Textes, Hammamet- Tunisie, 21-23 octobre 2002.

[3] [Beesley] Kenneth R. Beesley : Xerox Arabic Morphogical Analysis and Generation Romanization, Transcription and Transliteration. http://www.xrce.xerox.com/competencies/content-analysis/arabic/info/romanization.html

[4] [Buschbeck *et al*. 2002] Bushbeck Bianka, Grivel Luc, Guillemin-Lanne Sylvie, Lautier Christian « Une application industrielle d'extraction d'informations pour l'Intelligence Economique » EGC 2002 Extraction et Gestion des Connaissance, Montpellier, 21-23 janvier 2002.

[5] [[Chen et Gey, 2002] A. Chen and F. Gey : Building an Arabic Stemmer for Information Retrieval. *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*. National Institute of Standards andTechnology, Nov 18-22, 2002, *pp631-640*.

[6] [Coupet et Huot. 2005] Coupet Pascal, Huot Charles, « Le Text Mining sur la langue Arabe : application au traitement des sources ouvertes » Congrès SFBA 2005, 13-17 Juin 2005.Ile Rousse, France.

[7] [Debili, 2002] Debili F., Achour H., Soussi E. : La langue arabe et l'ordinateur : de l'étiquetage grammatical à la voyellation automatique, *Correspondances de l'IRMC, N° 71, juillet-août 2002, pp 10-28*

[8] [Debili, 2005] Debili F., Souissi E. (2005). Y a-t-il une taille optimale des règles de succession intervenant dans l'étiquetage grammatical ? Actes de *TALN'2005, Dourdan, Juin 2005*, 363-372.

[9] [Delecroix *et al* .2004] Delecroix Bertrand, Guillemin-Lanne Sylvie, Six Amandine « Veille concurrentielle et veille stratégique : deux applications d'extraction d'information » VSST 2004 Veille Scientifique et Stratégique, Toulouse, 25-29 oct 2004,pp 117-128.

[10] [Grivel *et al.* 2001] Grivel Luc, Guillemin-Lanne Sylvie, Coupet Pascal, Huot Charles « Analyse en ligne de l'information: une approche permettant l'extraction d'informations stratégiques basée sur la construction de composants de connaissance » VSST 2001 Veille Scientifique et Stratégique, Barcelone, 15-19 oct

[11] [Guillemin-Lanne et al. 2006] Guillemin-Lanne Sylvie, Six Amandine « La normalisation : nouveau challenge en extraction d'information » VSST 2006 Veille Scientifique et Stratégique, Lille, 16-17 janvier

[12] [Hobbs 1997] Hobbs J. R. et al. FASTUS : A Cascaded Finite-State Transducers for Extracting Information from Natural-Language Text. In E. Roche et Y. Schabes (eds.), *Finite-State Language Processin*g. Cambridge MA: MIT Press. (1997)

[13] [Kassas 2005] Dina EL KASSAS : Une étude contrastive de l'arabe et du français dans une perspective de génération multilingue http://www.olst.umontreal.ca/pdf/PhDElKassas2005.pdf

[14] [Neumann 1999] Neumann G., Schmeier S., Combining Shallow Text Processing and Machine Learning in Real World Applications. Proceedings of the IJCAI-99 workshop on Machine Learning for Information Filtering, Stockholm, Sweden, 1999.

[15] [Poibeau 2002] Poibeau T. : Extraction d'information à base de connaissances hybrides, Thèse, Université Paris-Nord, 8 mars 2002.

[16] [Wilks 97] Wilks, Y.Information Extraction as a Core Language Technology. In Pazienza, M. T., editor, *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, Frascati, Italy, LNAI Tutorial, Springer. pp. 14-18,1997.

[17] [Yangarger et Grishman, 1997] Yangarber R., Grishman R., “Customisation of Information Extraction Systems”. In Pazienza, M. T., editor, *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, Springer Verlag, Heildelberg, 1997, pp. 1-11.

[18] [Zanasi 2001] Zanasi, A. Text Mining: The New Competitive Intelligence Frontier. Real Application Cases in Industrial, Banking and Telecom/SMEs World» VSST 2001 Veille Scientifique et Stratégique, Barcelone, 15-19 octobre 2001.