Le traitement automatique de l'arabe, JEP-TALN 2004, Fès, 19-22 avril 2004

Filtrage sémantique de textes en arabe en vue d'un prototype de résumé automatique

Motasem ALRAHABI, Ghassan MOURAD, Brahim DJIOUA

LaLICC (Langage, Logique, Informatique, Cognition et Communication)
UMR 8139, Université Paris – Sorbonne, CNRS
96, Bd Raspail 75006 Paris – France
Tél. : (33) 01 44 39 35 90
[prenom.nom]@paris4.sorbonne.fr

Résumé – Abstract

Dans cet article, nous proposons un prototype de filtrage sémantique de textes en arabe, basé sur la méthode de l'exploration contextuelle. Son principe s'appuie sur des connaissances linguistiques et permet de repérer, grâce à des indices linguistiques, des informations pertinentes, comme les annonces thématiques, les énoncés définitoires, les titres, les soulignements, les récapitulations et les conclusions.

This paper describes a semantic filtiring system of texts in arabic language, based on the contextual exploration method. Its principle is based on linguistic knowledge and allows to find, with linguistic markers, relevant informations, like thematic segments, definition utterances, titles, underlinings, summing ups and conclusions.

Mots Clés – Keywords

Filtrage sémantique, résumé automatique, langue arabe, exploration contextuelle, marqueurs linguistiques.

Semantic filtiring, automatic summarization, arabic language, contextual exploration, linguistic markers.

1 Introduction

Le traitement automatique de la langue arabe est une discipline en pleine expansion, dans laquelle on voit de plus en plus de recherches et de technologies se soucier des spécificités de cette langue (Aloulou et al. 02, Baccour et al. 03, Boualem 93, Debili 01, Dichy et al. 02, Gaubert 01, Jaccarini 97, etc.) et proposer des outils nécessaires au développement de son traitement automatique.
Nous proposons dans cet article un système de filtrage sémantique de textes en arabe basé sur la méthode d'exploration contextuelle (Desclés 91, 97), un système qui fait appel à des connaissances strictement linguistiques et qui prend en compte, dans certains cas, les particularités de la langue arabe.
Nous savons que la notion de résumé automatique, facteur important de la gestion de l'information aujourd'hui, dépend complètement des attentes des utilisateurs d'une part, et de la nature et du type des documents traités d'autre part (Berri 96, Minel 02). Quelle information retenir dans un texte ? Peut-on a priori juger importante telle ou telle information à la place de l'utilisateur qui la cherche ?
Notre point de vue est qu'une information est plus ou moins importante selon ce que l'on cherche à savoir du texte, c'est pourquoi nous ne pouvons pas imaginer un « résumé idéal » (Desclés 03) qui serait indépendant des attentes des utilisateurs. Dans cette perspective, nous proposons une approche alternative aux méthodes classiques (Berri 96, Minel 02) adoptées dans ce domaine, il s'agit de la méthode d'exploration contextuelle (EC). Cette méthode permet d'accéder au contenu sémantique d'un texte sans avoir recours à des analyses syntaxiques profondes (Desclés 96) ou à des connaissances extérieures mais en prenant en compte un certain nombre de marqueurs linguistiques associés à une notion, ces marqueurs étant insérés dans un contexte où sont identifiés certains indices linguistiques.
Ainsi, en nous fixant comme tâche le repérage des informations résumantes dans un texte en arabe, nous nous voyons confrontés à plusieurs obstacles:

- Le problème d'encodage des textes traités ;
- Le choix de la structure logique des textes ;
- La définition des valeurs sémantiques associées à la fonction résumante ;
- La reconnaissance dans les textes des marqueurs linguistiques permettant d'identifier automatiquement cette notion ;
- Le décryptage du mécanisme régissant le fonctionnement de ces marqueurs ;
- Le choix de la technique informatique adéquate pour implémenter le système ;

Pour répondre à toutes ces questions nous montrons ci-dessous en détail la méthodologie avec laquelle nous avons réalisé ce projet.

2 Présentation du travail

La méthode d'EC est « capable d'identifier dans un texte certaines relations organisatrices de la connaissance ainsi que les organisations textuelles mises en place par l'auteur. » (Desclés 91). Ainsi lorsqu'un auteur s'exprime sur un sujet donné, il pourrait puiser dans ce que la langue lui offre d'éléments linguistiques, qui sont indépendants d'un domaine particulier et servent à modéliser les différentes manifestations textuelles. Ces éléments se présentent sous formes de régularités lexicales, grammaticales ou discursives,
En ce sens, le principe de notre système de résumé automatique est de systématiser cette approche en repérant dans les textes, au moyen de marqueurs linguistiques employés par l'auteur, les différentes informations que l'utilisateur recherche. Ensuite il s'agit, dans un formalisme de règles d'EC, de formuler le fonctionnement de ces éléments et de l'automatiser

de manière à ce qu'une machine puisse le faire ultérieurement sur des textes, et d'attribuer à certaines phrases des annotations sémantiques (Ben Hazez et al. 01) relatives aux informations identifiées. Ainsi, la mise en place de notre système s'articule sur deux axes: la gestion des connaissances linguistiques et l'implémentation informatique du système.

3 Gestion des connaissances linguistiques

La méthodologie suivie dans ce travail consiste en un ensemble d'étapes ordonnées avec d'éventuels retours en arrière : le choix du corpus de travail ; l'analyse linguistique de ce corpus, qui comprend essentiellement la recherche et l'organisation des marqueurs linguistiques ; et enfin l'exhibition de règles d'annotation de segments textuels ayant les valeurs sémantiques de résumé.

3.1 Constitution du corpus

C'est à partir du Web que nous avons construit notre corpus de textes en langue arabe.
Les textes, qui sont des articles de presse, ont été rapatriés sans restriction quant à leur nature ou leur volume. La raison en est que nous estimons que plus le corpus est étendu et varié, plus il sera représentatif et plus important sera le nombre de marqueurs linguistiques qu'il contiendra. Notre but est de construire un résumeur qui fonctionne avec tout type de texte, puisque le postulat de départ revient à considérer les segments textuels « filtrés » comme des expressions utilisées par l'auteur à l'aide de marqueurs indépendants du domaine.
La seule restriction a été le choix de textes en caractères non vocalisés. En fait, dans les éditions modernes, les textes en langue arabe ne sont pas souvent vocalisés. Nous rappelons que la vocalisation en arabe est l'ajout sur les consonnes de signes suscrits ou souscrits précisant la prononciation.
Le corpus final comporte environ cent textes dont la taille moyenne est de quatre pages. Il est suffisant à cette étape de recherche et sera incrémenté ultérieurement.

3.2 Analyse linguistique du corpus

La méthode d'EC, appliquée au filtrage sémantique de segments textuels, dépend de la notion sous-jacente à ce marquage sémantique. En ce qui concerne notre application, nous commençons d'abord par définir des valeurs sémantiques qui représentent le résumé et à choisir le genre d'informations à retenir dans un texte. Ces valeurs sémantiques (ou annotations) sont ensuite ordonnées dans un réseau organisé, une sorte de "carte sémantique", selon la conception de J.-P. Desclés. Nous travaillons pour le moment avec les étiquettes sémantiques suivantes : les titres, les énoncés définitoires, les annonces thématiques, les soulignements, les récapitulations et les conclusions.
Ensuite, nous commençons l'analyse sémantique du corpus par une lecture analytique et sélective afin de filtrer les segments textuels correspondant à ces étiquettes. Voici quelques exemples de phrases repérées dans le corpus :

سوف أنتقل الآن إلى صلب الموضوع... / *Je vais passer maintenant au cœur du sujet…*
كيف نواجه ظاهرة التصحر؟... / *Comment faire face à la désertification ?...*
من الجدير بالذكر أن هذا المرض... / *Il est intéressant de souligner que cette maladie…*
تعريف: إن مرض الإيدز هو... / *Définition : la maladie du Sida est…*
مما سبق نستطيع القول أن... / *De ce qui précède nous pouvons dire que…*
خاتمة: لقد كان هذا الشاعر... / *Conclusion : ce poète fut…*

La première phrase, par exemple, représente une annonce thématique. Celle-ci a généralement pour rôle d'expliciter le thème du document ou le plan suivi par l'auteur au fil du texte (Minel et al. 01). Nous allons isoler dans cette phrase les termes neutres, extérieurs au sujet traité et qui sont les marqueurs linguistiques représentant cette annonce thématique :

الآن / *maintenant*; الموضوع / *le sujet*; صلب / *cœur*; أننقل / *passer*.

Parmi ces marqueurs, nous devons distinguer entre les indicateurs linguistiques pertinents et les indices linguistiques complémentaires. Nous remarquons que le mot الموضوع / *sujet* est l'indicateur le plus important de cette annonce thématique, alors que le mot أننقل / *passer* devient un indice complémentaire.
Signalons ici que certains marqueurs peuvent remplir le rôle d'indicateurs pertinents dans certaines expressions alors qu'ils sont de simples indices complémentaires dans d'autres.

3.3 Organisation des marqueurs linguistiques en classes

Une fois que ces unités linguistiques ont été repérées, il s'agit de les stocker dans des listes distinctes qui seront incrémentées au fur et à mesure. Ainsi nous aurons, dans chaque liste (ou classe), une sorte de paradigme d'unités linguistiques dont les catégories sont parfois hétérogènes (noms, verbes, mots outils ou grammaticaux, etc.) mais qui remplissent toujours les mêmes fonctions sémantiques discursives. Voici quelques exemples de classes d'indicateurs :

Classe *Thème* : مشكلة / *problème*, قضية / *affaire*, مسألة / *question*, نقاش / *discussion*, etc.
Classe *Document* : بحث / *recherche*, كتاب / *livre*, مقالة / *article*, نص / *texte*, etc.
Classe *Explication* : أشار/ *signaler*, إيضاح/*éclaircissement*, تحدث/*discuter*, عالج / *traiter*, etc.
Classe *Définition* : تعريف / *définition*, يعرف / *connu*, يسمى / *appelé*, معناه / *signifiant*, etc.
Classe *Récapitulation* : بعبارة أخرى / *en d'autres termes*, بالمختصر المفيد / *en résumé*, etc.
Classe *Conclusion* : خاتمة /*Conclusion*, في النهاية / *à la fin*, etc.
Classe *Soulignement* : على الأخص/ *particulièrement*, ركز/ *insister*, خصوصاً / *surtout*, etc.

Et voici quelques classes d'indices complémentaires :

Classe *Auteur* : مؤلف/ *auteur*, حرر / *rédiger*, كاتب / *écrivain*, مترجم / *traducteur*, etc.
Classe *Transition* : تطرق/*toucher*, دخل/*entrer*, انتقال/*passage*, بدأ /*commencer*, باشر /*attaquer*, etc.
Classe *Subordonnant* : أن / *que*, إن/ *que*, etc.
Classe *Démonstratif* : هؤلاء / *ceux*, هذا/ *ce*, هذه / *cette*, ذلك/ *celui-là*, etc.

Le nombre de classe et de sous-classes s'élève actuellement à quarante.

3.4 Enrichissement des classes de marqueurs

Vu la particularité de la morphologie arabe au niveau de la vocalisation et d'agglutination, (88 علي ,قدور99, Boualem 93, Debili 01, Dichy et al. 02, Gaubert 01, Hakkak et al. 96, Jacarini 97), nous avons procédé à l'élaboration, pour les marqueurs collectés, de toutes les formes agglutinées ainsi que les autres variantes morphologiques susceptibles d'être rencontrées dans les textes. Ainsi, on a effectué sur la plupart des marqueurs, selon la nécessité du contexte, les opérations suivantes :

- La synonymie : pour دخل / *entrer* par exemple[1], on aura : تخلل / تغلغل / نفذ / ولج / توغل ;

[1] D'après le dictionnaire des synonymes : قاموس المترادفات والمتجانسات (1986), دار المشرق, بيروت

- La reformulation, la paraphrase ou la métaphorisation : pour le mot دخل / *entrer*, on aura des mots comme: تطرق, باشر, انتقل, بدأ, عالج, محص ;
- La flexion : nous avons ajouté pour certains mots un paradigme flexionnel selon :
  1. le genre : مهم/مهمة (*important/importante*) ;
  2. le nombre : نقاش/نقاشات (*débat/débats*) ;
  3. la conjugaison : شرح/شرحنا (*expliquer/nous avons expliqué*) ;
- La dérivation : nous avons ajouté une famille dérivationnelle pour certains termes : pour le mot كتب/ *a écrit*, par exemple, nous pouvons retrouver des mots tels : كاتب/écrivain, مكتوب/écrit, كتابة/écriture ;
- L'ajout de certaines formes agglutinées : il y a en arabe certains articles, pronoms, prépositions, conjonctions et autres particules qui se lient morphologiquement aux formes simples pour constituer un seul mot. La soudure de ces formes agglutinées rend difficile (Debili 01) la reconnaissance des marqueurs linguistiques sous leurs différentes formes. Ainsi nous avons rajouté à la base de données certaines formes agglutinées importantes:
  1. avec proclitiques : بالمقالة (*dans l'article*) ;
  2. avec enclitiques : مقالتي (*mon article*) ;

Cet enrichissement augmente considérablement la combinatoire des marqueurs collectés, dont la couverture n'a pas encore atteint, jusqu'à présent, un seuil de stabilité où l'on ne trouverait plus de nouveaux marqueurs pour la tâche donnée. Néanmoins, sur les six catégories discursives nous avons obtenu environ trois mille marqueurs linguistiques (indicateurs et indices).

3.5 Ecriture des règles d'EC

Les règles d'EC sont des heuristiques ; elles retracent fidèlement l'apparition dans les textes des marqueurs linguistiques en spécifiant leurs dépendances contextuelles. Elles sont conçues de manière à rechercher, dans un premier temps, un indicateur linguistique pertinent pour la tâche de résumé automatique, et ensuite à rechercher des indices complémentaires dans un espace de recherche défini à partir de cet indicateur. Si les conditions prédéterminées sont requises, une décision est alors prise : une annotation sémantique est attribuée à la phrase en question. Regardons l'exemple suivant :

من الجدير بالذكر أن هذا المرض يصيب الأطفال اعتباراً من سن السادسة

Il est intéressant de souligner que cette maladie touche les enfants à partir de six ans

Après analyse de cette phrase, nous pouvons dire : « Si nous trouvons dans le texte l'indicateur ذكر / *souligner* et si à droite de cet indicateur, dans la même phrase, nous avons l'indice من الجدير/ *il est intéressant*, nous considérons alors que la phrase en question indique un *Soulignement*. »

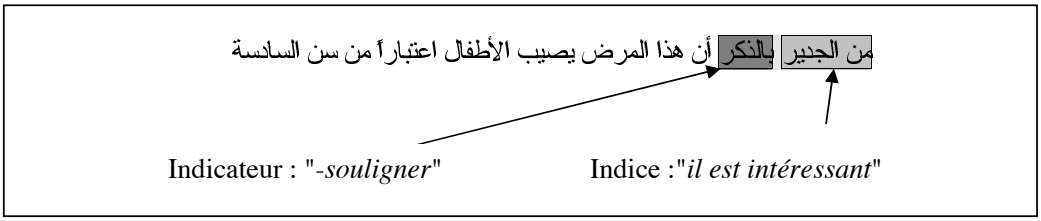

Figure 1 : Repérage d'indicateurs et d'indices dans un texte en arabe

Les règles d'EC sont exprimées dans un langage déclaratif simple, d'une manière indépendante de toute implémentation informatique. Nous avons construit une trentaine de règles pour les différentes catégories sémantiques de filtrage.

4 Réalisation informatique du système

L'implémentation de notre système dans un prototype informatique opérationnel trouve sa réalisation dans un outil automatique d'annotation sémantique. Ce système, spécialement créé pour cette tâche, est conçu en plusieurs couches (voir figure 3) : un langage d'expression d'automates à états finis (Silberztein 93) sous forme d'expressions régulières avec une extension Unicode. Une deuxième couche permet l'expression et l'interprétation de classes des marqueurs linguistiques ainsi que certains opérations de quantification (*, ?, +) sur ces classes. La dernière couche de ce langage permet l'expression de règles d'annotation sémantique sous forme de règles d'exploration contextuelle. L'application de celles-ci s'appuie d'abord sur le repérage d'un indicateur fort, porteur de la notion sémantique, ensuite, elle est confirmée ou non par la présence des indices complémentaires dans son contexte gauche et/ou droit, comme il a été bien indiqué plus haut.

4.1 Codage de l'écriture arabe

Il est clair que d'un point de vue informatique, un texte est une chaîne de caractères codée avec un système de codage quelconque comme ASCII, ISO-8859-n, Windows-125n, ASMO708, etc.

Les textes traités par notre système utilisent l'encodage UTF-8 (*Unicode Transformation Format*). Nous signalons à ce stade qu'avant l'avènement du standard Unicode, l'informatisation de l'écriture arabe (01خضر, Zaghibi 02) posait de sérieux problèmes techniques avec l'affichage de l'arabe sur différents systèmes d'exploitation, la bidirectionnalité, la nature cursive de l'écriture arabe et les signes de vocalisation. Ainsi, en spécifiant un numéro unique pour chaque caractère, Unicode a remédié à la majorité de ces difficultés.

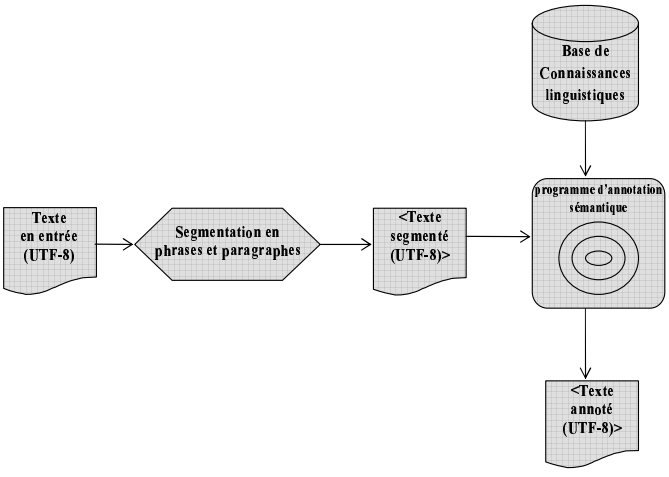

Figure 2 : Architecture générale du système d'annotation sémantique

4.2 Segmentation du corpus

Il s'agit, à cette étape, de segmenter les textes afin de déterminer les unités de traitement (phrase, paragraphe, section, etc.) dans lesquels le moteur va rechercher les marqueurs linguistiques. Pour notre corpus, nous avons utilisé un segmenteur pour la langue arabe mis au point par (Baccour et al. 03) du laboratoire LARIS. Ce segmenteur, basé sur l'EC, prend en entrée un fichier en format texte et fournit en sortie un fichier en format XML suivant une DTD bien définie (Pierel 00), avec des balises encadrant les phrases <ج> et les paragraphes <ف>.
Le principe de base pour trouver la fin des phrases (Baccour et al. 03) est d'étudier les contextes gauches et droits des éléments suivants : les signes de ponctuation (le point, la virgule, le point d'exclamation, les deux points), les conjonctions de coordination (le waw/ و, le thomma /ثم) et les mots connecteurs (/حتى même, /لكن mais, /لكنْ mais).
Signalons au passage une grande difficulté à ce niveau de traitement : à la différence des langues latines, la segmentation des textes en arabe ne peut pas s'appuyer sur les signes typographiques et les majuscules (Mourad 02).

4.3 Implémentation des marqueurs et des règles

Les marqueurs linguistiques sous forme de listes et les règles d'annotation sont stockés dans de simples fichiers avec un codage en UTF-8. Chaque Classe de marqueur est identifiée par le nom du fichier dans lequel elle est enregistrée.

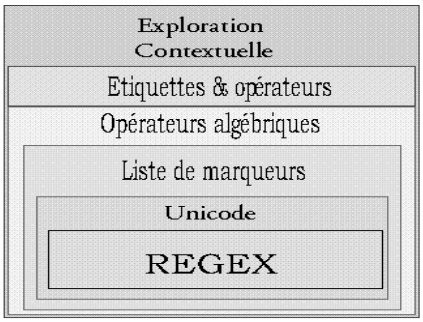

Figure 3 : Forme de prémisses des règles d'annotation

Les règles d'annotation sémantique sont exprimées avec un langage multi-couches. La forme générale d'une règle d'annotation est :

```
regle-s := regex | (1)
         liste-marqueur| (2)
         etiquette| (3)
         (liste-marqueur|etiquette|regex)+ (4)
regle:=([indicateur|indice]:regle-s\s)+\t[phrase|paragraphe]:etiquette (5)
liste-marqueur := (@liste[?|+|*]\s)+\tetiquette
etiquette := (<etiq>[ ?|+|*]\s)+\tetiquette
```

(1) : les prémisses de la règle sont exprimées sous forme d'expressions régulières, l'annotation portant sur la partie reconnue,

(2) : les prémisses de la règle sont exprimées à l'aide de listes de marqueurs et d'éventuels quantificateurs sur ses listes,

(3) : dans une prémisse, on peut exprimer la présence d'un segment textuel déjà annoté par une règle antérieure,

(4) : une prémisse peut mélanger les trois premières formes d'expression de prémisse,

(5) : la dernière forme des conditions d'application d'une annotation s'exprime par un typage de ses parties. Les types indice/indicateur permettent d'indiquer la manière d'identifier un segment textuel à annoter selon la méthode d'exploration contextuelle.

Prenons une règle d'EC qui exprime une annotation d'une annonce thématique dans un segment textuel « phrase » (<ج> ...</ج>), qui sera indiqué dans l'action de la règle par le type du segment textuel à annoter:

```
indicateur:@Particule-Thematique indice:@Soulignement-Thematique  phrase:Annonce Thématique
```

Appliquée à un texte du corpus, cette règle va engendrer des balises sémantiques de type :

<ج نمط="دلالة_مبحثية"><أما بالنسبة للتذييلات فإن الاستنساخ محصور بالقسام البيضة وهو ما يندرج تحت اسم التوأم وحيد المشيج.</ج>

La balise <ج> encadre une phrase. L'attribut qui vient après (نمط /étiquette) représente ici l'étiquette sémantique de la phrase soulignée : دلالة_مبحثية/*Annonce Thématique*.

5 Tests

Après avoir choisi le texte à résumer, l'utilisateur peut définir un profil de filtrage parmi les catégories discursives proposées. Le programme segmente le texte, y applique les règles d'EC concernées afin de rechercher le type d'informations désiré. La sortie est un texte décoré de balises sémantiques quand elles sont présentes.

Figure 4 : Un exemple de fonctionnement du système de résumé

Cette étape consiste à tester l'aspect opérationnel du système informatique. Elle est très importante dans la réalisation du projet car elle nous sert à faire des retours en arrière afin de vérifier la pertinence et la couverture des règles d'EC, et de faire les modifications nécessaires dans leur système d'inférence ainsi que sur le choix et l'organisation des marqueurs linguistiques. Les tests montrent certaines limites de notre système de filtrage, au niveau de la reconnaissance des marqueurs dans les textes et de la capacité des règles d'EC à couvrir toutes les formes textuelles possibles. Ce constat nous incite à chercher de nouveaux moyens d'amélioration que nous proposons dans la suite du document.
A ce stade du développement du prototype, nous ne pouvons pas encore fournir une évaluation précise avec une procédure d'annotation spécialement adaptée à l'activité résumante (Minel et al. 01).

6 Conclusion et perspectives

Le système de résumé automatique des textes en arabe s'inscrit dans le cadre du filtrage sémantique des textes à l'aide de critères purement linguistiques et trouve sa réalisation avec la méthode d'exploration contextuelle. Cette méthode s'avère parfaitement adaptable à la langue arabe et aux particularités de sa forme textuelle ; elle nous permet de repérer des phrases contenant des informations recherchées par l'utilisateur, comme les annonces thématiques, les titres, les expressions définitoires, les récapitulations, les soulignements, les conclusions.
Dans le cadre de ce projet nous étudions des perspectives pouvant améliorer le fonctionnement de notre application.
Au niveau linguistique, un travail d'analyse plus approfondie de textes reste à faire. Ce qui nous permettra d'écrire des règles d'EC plus précises et en plus grand nombre, capables de couvrir plus d'expressions discursives. Cela permettra en même temps d'enrichir la base de données des marqueurs linguistiques. Nous pouvons aussi ajouter de nouvelles étiquettes sémantiques discursives comme l'identification des citations.
Au niveau informatique, un travail de développement des fonctionnalité du programme est en cours. Cela nous permettra par conséquent d'effectuer les tests et les évaluations des résultats du système. Nous étudions la possibilité d'intégrer dans notre système un analyseur morphosyntaxique de l'arabe (Gaubert 01, Jaccarini 97) ou un dictionnaire (Dichy et al. 02) permettant la désambiguïsation des différentes formes des marqueurs linguistiques utilisés dans l'EC.

Remerciements

Nous tenons à remercier Monsieur le professeur Jean-Pierre Desclés pour ses conseils précieux et ses encouragements.

Références

Aloulou C., Belguith Hadrich L. et Ben Hamadou A. (2002), *—Utilisation des grammaires HPSG pour l'analyse de l'Arabe*, JEI'2002, 2ème journée des jeunes chercheurs en électronique et Informatique, Hammamet, Tunisie.
Baccour L., Mourad G., Belguith Hadrich L. (2003), *Segmentation de textes arabes en phrases basée sur les signes de ponctuation et les mots connecteurs*, troisième journées scientifiques des jeunes chercheurs en génie électrique et informatique, du 25-27 mars, Mahdia, Tunisie.

Ben Hazez S., Desclés J.-P., Minel J.-L., *Modèle d'exploration contextuelle pour l'analyse sémantique de textes*", TALN 2001, 2001.
Berri J. (1996), *Contribution à la méthode d'Exploration Contextuelle. Applications au résumé automatique et aux représentations temporelles. Réalisation informatique du système SERAPHIN*, Thèse de doctorat, Université de ParisIV-Sorbonne, Paris.
Boualem M. (1993), *Système de conversion de formalismes de langages techniques dans un environnement à syntaxe contrôlée et à contexte limité*, Thèse de doctorat, Université de Sophia Antipolis.
Dichy J., Braham A., Ghazali S., Hassoun M. (2002), La base de connaissances linguistiques DIINAR.1, DIctionnaire INformatisé de l'ARabe, Actes du colloque *Proceeding of the International Sympotium*, université de Manouba, Tunisie.
Debili F. (2001), Traitement automatique de l'arabe voyellé ou non, *Correspondances* n°46, IRMC, Tunis.
Desclés et al. (1991), Exploration Contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte. In *Knowledge modeling andexpertise transfer*, pp.371-400, D. Herin-Aime, R. Dieng, J-P. Regourd, J.P. Angoujard (éds), Amsterdam.
Desclés J.-P. et al. (1996), Filtrage automatique des textes, *NLP&IA International Conference on Natural Language Processing and Industrial Applications*, Moncton, Canada.
Desclés J.-P. (1997), *Systèmes d'Exploration Contextuelle. Co-texte et calcul du sens*, (ed. Claude Guimier), Presses Universitaires de Caen, p. 215-232.
Desclés J.-P. (2003), *Ingénierie linguistique : enjeux, domaines et méthodes*, conférence université Klément d'Okrid, Sofia.
Gaubert C. (2001), *Stratégies et règles minimales pour un traitement automatique de l'arabe*, Thèse de doctorat, Université Aix-Marseille I.
Hakkak G. et Neyreneuf M. (1996), *Grammaire active de l'arabe*, Librairie Générale Française.
Jaccarini A. (1997), *Grammaire modulaires de l'arabe. Modélisation, mise en oeuvre informatique et stratégies*. Thèse de doctorat, Université de ParisIV-Sorbonne, Paris
Minel J.-L. (2002), *Filtrage sémantique, du résumé automatique à la fouille de textes*, Hermès, Paris.
Minel J.-L., Desclés J.-P., Cartier E., Crispino G., Ben Hazez S., Jackiewicz A., (2001), Résumé automatique par filtrage sémantique d'informations dans des textes, Technique et Science Informatiques.
Mourad G. (2002), La segmentation de textes par Exploration Contextuelle automatique, présentation du module SegATex, Inscription Spatiale du Langage : structure et processus *ISLsp*., Toulouse.
Pierel J.-M, sous la direction de, *Ingénierie des langues* (2000), Hermès, Paris, 2000.
Silberztein M. (1993), *Dictionnaires électroniques et analyse automatique de textes, le système INTEX*, Masson, Paris.
Zaghibi R. (2002), Le codage informatique de l'écriture arabe : d'ASMO449 à UNICODE et ISO/CEI10646, in *Unicode, écriture du monde*, Hermes.

Références en arabe :

خضر, محمد زكي (2001), الحروف العربية والحاسوب, مجلة مجمع اللغة العربية, عمان
علي, نبيل (1988), اللغة العربية والحاسوب, دراسة بحثية, شركة العريض
قدور, أحمد محمد (1999), مبادئ اللسانيات, دار الفكر, دمشق