Résumés de thèses

Rubrique préparée par Fiammetta Namer

Université Nancy2 de Nancy, UMR « ATILF »
Fiammetta.Namer@univ-nancy2.fr

Nicolas BÉCHET (bechet@lirmm.fr**)**

Titre : Extraction et regroupement de descripteurs morpho-syntaxiques pour des processus de Fouille de Textes.

Mots-clés : TAL, fouille de textes, descripteur, syntaxe, classification.

Title : *Extraction and Gathering of morpho-syntactic features for Text Mining process.*

Keywords : *NLP, text mining, feature, syntax, categorization.*

Thèse de doctorat en , Université Montpellier 2, UFR Sciences et Techniquessous la direction de Jacques Chauché (Pr, Université de Montpellier 2) et Mathieu Roche (MC, Université de Montpellier 2). Thèse soutenue le 08/12/2009.

Jury : M. Jacques Chauché (Pr, Université de Montpellier2, directeur), M. Mathieu Roche (MC, Université de Montpellier 2, codirecteur), Mme Catherine Berrut (Pr, Université Joseph Fourier, rapporteur), M. Christophe Roche (Pr, Université de Savoie, rapporteur), Mme Violaine Prince (Pr, Université de Montpellier 2, examinatrice), Mme Anne Vilnat (Pr, Université de Paris-Sud, examinatrice).

Résumé *: Les mots constituent l'un des fondements des langues naturelles de type indo-européen. Des corpus rédigés avec ces langues sont alors naturellement décrits avec des mots. Cependant, l'information qu'ils véhiculent seuls est assez réduite d'un point de vue sémantique. Il est en effet primordial de prendre en compte la complexité de ces langues comme, par exemple, leurs propriétés syntaxiques, lexicales et sémantiques. Nous proposons dans cette thèse de prendre en considération ces propriétés en décrivant un corpus par le biais d'informations syntaxiques permettant de découvrir des connaissances sémantiques.*

Nous présentons dans un premier temps un modèle de sélection de descripteurs SelDe. Ce dernier se fonde sur les objets issus des relations syntaxiques d'un corpus. Le modèle SelDe a été évalué pour des tâches de classification de données

textuelles. Pour cela, nous présentons une approche d'expansion de corpus, nommée ExpLSA, dont l'objectif est de combiner les informations syntaxiques fournies par SelDe et la méthode numérique LSA (« latent semantic analysis »).

Le modèle SelDe, bien que fournissant des descripteurs de bonne qualité, ne peut être appliqué avec tous types de données textuelles. Ainsi, nous décrivons dans cette thèse un ensemble d'approches adaptées aux données textuelles dites complexes. Nous étudions la qualité de ces méthodes avec des données syntaxiquement mal formulées et mal orthographiées, des données bruitées ou incomplètes et finalement des données dépourvues de syntaxe.

Finalement un autre modèle de sélection de descripteurs, nommé SelDeF, est proposé. Ce dernier permet de valider de manière automatique des relations syntaxiques dites « induites ». Notre approche consiste à combiner deux méthodes. Une première approche fondée sur des vecteurs sémantiques utilise les ressources d'un thésaurus. Une seconde s'appuie sur les connaissances du Web et des mesures statistiques afin de valider les relations syntaxiques. Nous avons expérimenté SelDeF pour une tâche de construction et d'enrichissement de classes conceptuelles. Les résultats expérimentaux montrent la qualité des approches de validation et reflètent ainsi la qualité des classes conceptuelles construites.

URL où la thèse pourra être téléchargée :
http://www.lirmm.fr/~bechet/These/These.pdf

Mohsen MARAOUI (maraoui.mohsen@gmail.com)

Titre : Élaboration d'un dictionnaire multifonction, à large couverture, de la langue arabe. Applications aux systèmes d'ALAO.

Mots-clés : TAL, ALAO, dictionnaire multifonction, langue arabe.

Title : *Development of a multifunctional dictionary, with broad coverage for the Arabic language. Applications to CALL systems.*

Keywords : *NLP, CALL, Multifunction Dictionary, Arabic Language.*

Thèse de doctorat en Sciences du langage, spécialité : Industries de la Langue, Département Informatique pédagogique et UMR LILIDEM, Université Grenoble 3-Stendhal, Grenoblesous la direction de Georges Antoniadis (MC-HDR, Université Grenoble 3). Thèse soutenue le 18/12/2009.

Jury : M. Georges Antoniadis (MC-HDR, Université Grenoble 3, directeur), M. Ahmed Jerraya (DR, CEA-LETI, président), M. Jean Caelen (DR, LIG-CNRS, rapporteur), M. Mounir Zrigui (MC-HDR, Université de Monastir, Tunisie,

rapporteur).

Résumé *: Il existe beaucoup de logiciels d'ALAO sur Internet. Il s'agit le plus souvent d'exercices à trous ou de QCM, conçus à l'aide de systèmes auteurs comme Course builder, Hot Potatoes ou Netquizz. Ce type d'activités pose plusieurs problèmes comme la rigidité des logiciels (les données utilisées sont prédéfinies et ne peuvent être ni modifiées ni enrichies) et la non-adaptabilité des parcours aux compétences linguistiques des apprenants (le cheminement est indépendant de ses réponses à chaque étape, faute de pouvoir les évaluer). L'unique avantage de ce genre d'exercices par apport à ceux rédigés sur papier est l'interactivité que permet l'ordinateur.*
L'utilisation du TAL, pour la conception de logiciels d'ALAO, représente un moyen faible pour résoudre ces problèmes.

Après plus de deux décennies, depuis le début des travaux, l'avancée des recherches dans ce domaine (c'est-à-dire l'utilisation de TAL en ALAO) reste insuffisante à cause principalement de deux facteurs : la méconnaissance du TAL de la part des didacticiens des langues, voire des informaticiens, et le coût des ressources et produits issus du traitement automatique de la langue. Il n'existe en effet pratiquement aucun produit commercialisé, à part un certain nombre de prototypes ou de systèmes expérimentaux pour les langues latines. Les travaux d'ALAO fondé sur le TAL pour la langue arabe sont pratiquement inexistants, malgré une riche bibliographie concernant le traitement automatique de l'arabe. Outre les facteurs cités précédemment, la carence concernant la langue arabe dans ce domaine est due au fait qu'elle est une langue difficile à traiter automatiquement.

En fonction de cette situation, et désireux d'enrichir les possibilités de création d'activités pédagogiques pour l'arabe, nous avons conçu un étiqueteur, un dérivateur, un conjugueur, un analyseur morphologique ainsi qu'un dictionnaire étiqueté (le plus complet possible) pour la langue arabe. Ensuite, nous avons exploité ces outils pour créer bon nombre d'applications pédagogiques pour l'apprentissage de l'arabe.

URL où la thèse pourra être téléchargée :
s'adresser à l'auteur

(*Erratum* : dans le TAL n°50-1/2009 nous avions indiqué que le lieu de soutenance de la thèse qui suit était Montréal, or il s'agissait de Lyon.)

Zoubeir MOUELHI (zoubeir.mouelhi@univ-lyon2.fr**)**

Titre : Essai de lexicométrie d'une œuvre arabe classique : *Al-'Imtâ' wa-l-Mu'ânasa* de Tawhîdî

Mots-clés : lexicométrie arabe, norme lexicologique, dépouillement lexical, TAL arabe, segmentation, lemmatisation, désambiguïsation, catégorisation, structure lexicale, richesse lexicale, catégories lexicales, textes arabes classiques, linguistique de corpus.

Title : *A lexicometrical study of an Arabic classic: Al-'Imtâ' wa-l-Mu'ânasa by Tawhîdî.*

Keywords : *Arabic lexicometry, lexicological norm, lexical data retrieval, Arabic NLP, segmentation, lemmatization, disambiguation, categorisation, lexical structure, lexical richness, lexical categories, classic Arabic texts, corpus linguistics.*

Thèse de doctorat en linguistique, Université de Lyon 2, département d'études arabes/Faculté des langues, laboratoire ICAR, Lyon, sous la direction de M. Joseph Dichy, (Professeur). Thèse soutenue le 22/11/2008.

Jury : M. Joseph Dichy (Pr, Université Lumière-Lyon 2, directeur), M. Hassan Sahloul (Pr, Université Jean Moulin-Lyon 3, rapporteur), M. Abdelfattah Braham (Pr, Université la Manouba, Tunisie, rapporteur), M. Xavier Lelubre (MCF-HDR, Université Lumière-Lyon2, examinateur), M. Mohamed Hassoun (Pr, ENSSIB, Villeurbane, examinateur).

Résumé *: S'inscrivant dans la perspective générale de l'approche quantitative de l'étude des textes, à l'intersection de plusieurs disciplines, notamment la linguistique, l'informatique et la statistique, l'approche lexicométrique trouve plusieurs applications eu égard aux textes, qu'ils soient pris isolément (préoccupations d'ordre stylistique, didactique, historique, etc.), comparés entre eux (typologies de textes, approche contrastive, etc.), considérés dans leur relation aux auteurs (homogénéité d'auteurs, attribution d'auteurs, etc.) ou dans leur relation au temps (séries textuelles chronologiques, spécificités chronologiques, etc.).*

Dans cette perspective lexicométrique, notre travail qui porte sur un ouvrage célèbre de la pensée arabe médiévale, l'Imtâ' wa-l-Mu'ânasa de Tawhîdî (IV^e^/X^e^ siècles) se fixe un triple objectif.

En premier lieu, l'élaboration, pour l'arabe, de ce que l'on appelle une norme lexicologique, donnant une assise théorique et méthodologique aux travaux lexicométriques futurs sur les textes arabes. Deux volets composent la norme lexicologique que nous proposons : une norme de saisie et d'harmonisation et une norme de dépouillement.

En deuxième lieu, la confection du dictionnaire de fréquences de notre corpus. Fruit naturel de toute étude lexicométrique globale de cette nature, le dictionnaire de fréquences traduit et synthétise les réorganisations formelles opérées sur la séquence textuelle d'origine, ainsi que le résultat des différentes analyses statistiques qui ont porté sur le vocabulaire du texte.

En troisième et dernier lieu, soumettre ce corpus à un certain nombre de méthodes d'analyse et de traitements statistiques propres à la lexicométrie en vue d'en étudier, principalement, la structure lexicale mais aussi la trame radicale. Ainsi, la richesse lexicale, l'accroissement du vocabulaire, la répartition des catégories lexicales, la connexion lexicale, etc., qui représentent tant d'éléments et d'indices pouvant caractériser le style d'un auteur, d'un genre ou d'une époque, ont-ils été étudiés et analysés. Il est nécessaire dans ce type d'entreprise, que des opérations de dépouillement préalables soient opérées selon des règles claires et stables assorties d'une réflexion minutieuse autour des notions de segmentation, de lemmatisation, de désambiguïsation, de catégorisation, etc. Les décomptes obtenus suite à ces étapes de dépouillement et de quantification, sont soumis aux traitements statistiques et à l'interprétation pour juger in fine *des variations des différentes unités linguistiques du corpus et en décrire la structure lexicale.*

URL où la thèse pourra être téléchargée :
http://theses.univ-lyon2.fr/

Hong Thai NGUYEN (Hong-Thai.Nguyen@imag.fr**)**

Titre : Des systèmes de TA homogènes aux systèmes de TAO hétérogènes.

Mots-clés : TAO hétérogène, TA, BDLM (base de données lexicale multilingue), re-ingénierie de LSPL (langage spécialisé pour la programmation linguistique), EDL (environnement de développement linguiciel).

Title : *From homogeneous MT systems to heterogeneous CAT systems*.

Keywords : *heterogeneous CAT, MT, lexical database, SLLP reengineering, EDL (Environment for Lingware Development).*

Thèse de doctorat en Informatique, École Doctorale « Mathématiques, Sciences et Technologie de l'Information », Université Joseph Fourier, laboratoire LIG, équipe GETALP, sous la direction de Christian Boitet (Pr, Université Joseph Fourier) et de Eric Castelli (MC HDR, INP Grenoble). Thèse soutenue le 18/12/2009.

Jury : M. Christian Boitet (Pr, Université Joseph Fourier, directeur), M. Eric Castelli (MC HDR, INP Grenoble, codirecteur), M. Laurent Besacier (Pr, Université Joseph Fourier, président), M. Jacques Chauché (Pr, Université Montpellier 2, rapporteur), M. Denis Maurel (Pr, Université de Tours, rapporteur), M. Jesus Cardeñosa (Pr, Université de Madrid, rapporteur), M. Vincent Berment(Ing., INALCO, examinateur), M. Mathieu Lafourcade (MC, Université Montpellier 2, examinateur).

Résumé *: Cette thèse porte sur les problèmes posés par la conception et la réalisation de la partie logicielle des systèmes de traduction automatisée (TAO) hétérogènes, intégrant des systèmes de TA multiples et/ou à composants hétérogènes, ainsi qu'une partie THAM (traduction humaine aidée par la machine), reposant sur des mémoires de traductions. Ces systèmes se développent à côté des systèmes de TA homogènes et de THAM, et les supplanteront peut-être à moyen terme.*
Leurs différents composants de TA seront construits par des équipes différentes, distribuées autour de la planète, avec des méthodes algorithmiques et des outils différents (langages spécialisés ou LSPL), ainsi que des ressources et composants linguiciels également différents (dictionnaires et corpus de divers types, grammaires et transducteurs fondés sur des règles), à l'aide d'EDL (environnements de développement linguiciel) eux aussi différents. Les contributions de la thèse concernent en particulier :
– l'amélioration des « méta-EDL de TAO », permettant d'effectuer une transition incrémentale entre les EDL natifs des systèmes de TA utilisés pour construire un systèmes de TAO à composants hétérogènes, et un futur EDL intégrateur universel, dans lequel on pourra « rapatrier » la compilation et l'exécution des LSPL ;
– la conception et la réalisation d'une base lexicale partageant un même pivot lexical, PIVAX, réalisée au-dessus de la plate-forme Jibiki (G. Sérasset, GETALP) ;
– la réingénierie de langages spécialisés « externes » (non supportés par l'EDL Ariane-G5), avec application aux « systèmes-Q » (A. Colmerauer, 1967), qui servit de base pendant 15 ans au système de TA TAUM-météo destiné aux bulletins météorologiques canadiens ;
– la conception et la réalisation d'un « moniteur » adapté à la partie « production » d'un système de TAO hétérogène, EMEU_w.1.0, qui a été développé et utilisé dans le cadre d'un projet de grande ampleur.

URL où la thèse pourra être téléchargée :
http://www-clips.imag.fr/geta/hong-thai.nguyen/these/these.pdf

---

Emmanuel PROCHASSON (eprochasson@gmail.com)

Titre : Alignement multilingue en corpus comparables spécialisés.

Mots-clés : corpus comparables, langue de spécialité, alignement multilingue.

Title : *Multilingual alignment in specialised comparable corpora.*

Keywords : *comparable corpora, specialised language, multilingual aligment.*

Thèse de doctorat en Informatique, UFR des Sciences et Techniques, Université de Nantes, UMR 6241 LINAsous la codirection de Béatrice Daille (Pr, Université de

Nantes) et d'Emmanuel Morin (Pr, Université de Nantes). Thèse soutenue le 17/12/2009.

Jury : Mme Béatrice Daille (Pr, Université de Nantes, directrice), M. Emmanuel Morin (Pr, Université de Nantes, codirecteur), M. Kamel Smaïli (Pr, Université Nancy2, président), M. Eric Gaussier (Pr, Université Joseph Fourier, rapporteur), M. Yves Lepage (Pr, Université de Caen, rapporteur).

***Résumé** : Les corpus comparables sont l'objet d'une attention particulière depuis le milieu des années 1990. Ils sont constitués de documents dans des langues différentes n'étant pas en relation de traduction, à l'inverse des corpus parallèles. Ils sont donc plus faciles à construire que ces derniers car moins contraints. En effet, les corpus parallèles nécessitent un travail coûteux de traduction humaine qui limite généralement leur disponibilité aux textes légaux (dépôt de brevet, actes de parlements...) ou à des traductions de ou vers l'anglais. À l'inverse, les corpus comparables sont généralement plus disponibles, en tout cas plus faciles à constituer dans de nombreuses langues et pour des thématiques variées. Ce travail porte sur l'extraction lexicale bilingue à partir de corpus comparables. Il s'agit du processus consistant à trouver un vocabulaire commun dans les différentes sous-parties des corpus, puis à l'aligner dans le but d'obtenir des paires de traductions. L'objectif du processus d'extraction et d'alignement de traductions est de permettre de compléter automatiquement des ressources lexicales multilingues. Ces ressources sont précieuses par exemple pour la traduction statistique automatique, mais aussi pour assister le travail des lexicographes et des terminologues en réalisant une partie du travail préliminaire de reconnaissance. La méthode que nous utilisons pour aligner des paires de traductions s'appuie sur la caractérisation du lexique dans un cadre monolingue.*

La caractérisation d'un mot – et a fortiori *d'un terme – est enregistrée sous la forme d'un vecteur de contexte. Ces vecteurs sont des structures de données contenant des informations sur l'environnement des mots. Ce sont ces informations que nous comparons entre les vecteurs d'une langue source et ceux d'une langue cible pour obtenir des candidats à la traduction. En effet, nous émettons l'hypothèse que les paires de traductions ont des environnements similaires, et donc des vecteurs de contexte similaires. Nous nous intéressons en particulier à la caractérisation terminologique car nous nous concentrons sur des textes en langue de spécialité. Nous travaillons sur deux corpus : le premier rassemble des documents en anglais et en français et traite du cancer du sein ; le second contient des documents en anglais, français et japonais et concerne la thématique diabète et alimentation, ce qui nous permet de traiter la question délicate de l'alignement du japonais et de langues indo-européennes. Ces corpus sont peu volumineux (quelques centaines de milliers de mots) : c'est une contrainte supplémentaire pour l'alignement que nos propositions doivent traiter en conséquence. Deux tâches nous intéressent particulièrement : la première est de définir quelles informations*

enregistrer effectivement dans les vecteurs de contexte ; la seconde est de savoir comment exploiter ces informations.

Une première approche se concentre sur la première tâche : nous essaierons de rendre les vecteurs de contexte plus discriminants, autrement dit, plus à même de stocker des informations nous permettant de trancher entre des éléments en relation de traduction et des éléments qui ne le sont pas. Nous réalisons cette tâche en nous appuyant sur une sous-partie connue du vocabulaire spécialisé : les points d'ancrage. Ces points d'ancrage seront employés pour leur fiabilité. Ils serviront à « déformer » les vecteurs de contexte des mots en leur accordant un plus grand poids pour l'étape de comparaison.

Une deuxième proposition se penchera sur la seconde tâche, en exploitant des ressources multisources pour renforcer la confiance dans un choix de traduction. Nous combinerons les informations apportées par des alignements anglais-japonais et français-japonais pour discriminer les traductions japonaises obtenues.

Une dernière proposition, relative aux deux tâches, étudiera l'importance de la fréquence des mots à traduire pour déterminer a priori *la meilleure façon de constituer les vecteurs de contexte dans le but d'obtenir un alignement optimal. Cette approche combine indirectement les informations apportées par les dépendances syntagmatiques, aptes à caractériser les mots fréquents, et les dépendances paradigmatiques, plus efficaces pour caractériser les mots de fréquences plus faibles.*

Nous prolongeons ce manuscrit par une réflexion sur la nature des corpus comparables et la notion de comparabilité.

URL où la thèse pourra être téléchargée :
http://eprochasson.free.fr/pperso/documents/these.pdf

Mathieu VALETTE (mvalette@atilf.fr)

Titre : Approche textuelle du lexique.

Mots-clés : lexique, sémantique des textes, dictionnaire sémique, sémimétrie, conceptualisation, néosémie.

Title : *Text-based approach to lexicon.*

Keywords : *lexicon, text-based semantics, seme dictionary, semimetry, conceptualization, neosemy.*

HDR en Sciences du langage, INALCO, EA 2520 CRIM-ERTIM, Parissous la direction de François Rastier (DR, CNRS). HDR soutenue le 08/12/2009.

Jury : M. François Rastier (DR, CNRS, directeur), M. André Salem (Pr, Université Paris 3, président), M. Salah Mejri (Pr, Université Paris 13, rapporteur), M. Pierre Zweigenbaum (DR, LIMSI-CNRS, rapporteur), Mme Annie Montaud (Pr, INALCO, examinatrice), Mme Sylviane Rémi-Giraud (Pr, Université de Lyon2, examinatrice).

Résumé *: La linguistique doit prendre part et position face aux nouveaux enjeux théoriques et méthodologiques naissant autour du document numérique et de l'élaboration des connaissances, et ne pas laisser à d'autres disciplines (sciences de l'information et de la communication, informatique) le soin de décrire, seules, ces nouveaux objets sémiotiques. Leur diversité et leur complexité sont en outre à problématiser tant dans la perspective de la variété des pratiques sociales que dans celle du multilinguisme. L'élaboration conjointe de modélisations linguistiques et d'outils informatiques destinés à leur validation et leur mise en œuvre s'avère une condition nécessaire à leur description. Dans ce cadre général, notre objectif est de présenter un ensemble de propositions visant à situer l'étude du lexique dans le paradigme textuel. Plus précisément, notre projet est d'étudier les déterminations textuelles de la conceptualisation et de la lexicalisation des concepts.*

Dans le premier chapitre, nous effectuons une revue critique des principaux modes de structuration et de représentation du lexique, en particulier dans la perspective d'un traitement automatique du sens. Nous exposons ensuite certaines propositions de la sémantique interprétative et textuelle de F. Rastier en la matière. Après une présentation de la notion de classes sémantiques, nous nous focalisons sur l'une d'entre elles, le taxème, et nous discutons plus particulièrement de son rôle dans la représentation de la praxis. Dans le deuxième chapitre, nous traitons de la représentation du lexique du point de vue du texte, c'est-à-dire du point de vue de l'agencement syntagmatique. Nous abordons les différentes objectivations sémantiques proposées par la théorie susmentionnée (isotopies, molécules sémiques) de façon à mettre en évidence le rôle de l'articulation lexique/texte dans la cohésion textuelle. Dans le troisième chapitre, nous présentons un ensemble de travaux réalisés dans la perspective d'une instrumentation de l'analyse sémantique des textes et du lexique faisant la synthèse des recherches relatées dans les deux précédents chapitres. Enfin, dans un quatrième chapitre, nous abordons la question de la conceptualisation et de la lexicalisation des concepts. Nous proposons une méthode de description fondée sur les propositions théoriques et les outils informatiques décrits précédemment. Nous présentons, finalement, un ensemble de prospectives et un programme de recherche relatif à l'approfondissement de notre approche dans la perspective des nouvelles applications de la linguistique, en particulier dans un contexte variationniste et multilingue.

URL où l'HDR pourra être téléchargée :
http://www.revue-texto.net