Résumés de thèses

Rubrique préparée par Fiammetta Namer

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

Al Moatasem ALRAHABI (motasem.alrahabi@gmail.com**)**

Titre : EXCOM-2 : plateforme d'annotation automatique de catégories sémantiques. Conception, modélisation et réalisation informatique. Application à la catégorisation des citations en français et en arabe.

Mots-clés : plateforme, annotation automatique, segmentation, exploration contextuelle, EXCOM2, analyse sémantique, marqueurs discursifs, carte sémantique, multilinguisme, discours rapporté.

Title : *EXCOM-2 : a platform for automatic annotation of semantic categories. Conception, modeling and implementation. Application to the categorisation of quotations in French and Arabic.*

Keywords : *Platform, automatic annotation, segmentation, Contextual Exploration, EXCOM2, semantic analysis, discoursive markers, Semantic Map, multilingual approach, reported speech.*

Thèse de doctorat en Mathématiques, Informatique et Application aux Sciences de l'Homme, Université Paris-Sorbonne UFR Sciences et TechniquesInstitut des Sciences Humaines Appliquées (ISHA), Laboratoire Langues, Logiques, Informatique, Cognition (LaLIC)sous la direction de Jean-Pierre Desclés (Pr, Université de Paris-Sorbonne). Thèse soutenue le 29/01/2010.

Jury : M. Jean-Pierre Desclés (Pr, Université de Paris-Sorbonne, directeur), M. Christian Boitet (Pr, Université Joseph Fourier, président), Mme Sabine Bergler (Associate Pr, Université de Concordia, Montréal, CA, rapporteur), M. Owen Rambow (Research Scientist, Université du Québec, Montréal, CA, rapporteur), M. JeanGuy Meunier (Pr, Université du Québec, Montréal, CA, examinateur), M. Christian Fluhr (Pr CEA, examinateur), M. Brahim Djioua (MC, Université de Paris-Sorbonne, examinateur).

Résumé *: Nous proposons une plateforme d'annotation sémantique, appelée « EXCOM2 ». Fondée sur la méthode de l'« exploration contextuelle », elle permet,*

à travers une diversité de langues, de procéder à des annotations automatiques de segments textuels par l'analyse des formes de surface dans leur contexte. Les textes sont traités selon des « points de vue » discursifs dont les valeurs sont organisées dans une « carte sémantique ». L'annotation se fonde sur un ensemble de règles linguistiques, écrites par un analyste, qui permet d'identifier les représentations textuelles sous-jacentes aux différentes catégories de la carte. Le système offre, à travers deux types d'interfaces (développeur ou utilisateur), une chaîne de traitements automatiques de textes qui comprend la segmentation, l'annotation et d'autres fonctionnalités de post-traitement. Les documents annotés peuvent être utilisés, par exemple, pour des systèmes de recherche d'information, de veille ou de résumé automatique. Comme exemple d'application, nous proposons un système d'identification et de catégorisation automatiques du discours rapporté direct en arabe et en français.

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

Delphine BATTISTELLI (delphine.batistelli@paris-sorbonne.fr)

Titre : La temporalité linguistique : circonscrire un objet d'analyse ainsi que des finalités à cette analyse.

Mots-clés : temps, aspect, modalité, énonciation, linguistique textuelle, linguistique énonciative, navigation textuelle, recherche d'information.

Title : *Linguistic Temporality. Defining the contours of an object of analysis and the purposes of such an analysis.*

Keywords : *tense, aspect, modality, enunciation, discourse linguistics, enunciative linguistics, textual navigation, information retrieval.*

Mémoire de HDR en Sciences du Langage (spécialité Traitement Automatique des Langues), Université Paris-Sorbonne, UFR Sciences et TechniquesInstitut des Sciences Humaines Appliquées (ISHA), Laboratoire EA STIH (Sens, Texte, Informatique, Histoire)sous la direction de Jean-Luc Minel (Pr, Université de Paris-Nanterre). Thèse soutenue le 23/11/2009 à l'Université Paris-Ouest Nanterre La Défense.

Jury : M. Jean-Luc Minel (Pr, Université de Paris-Nanterre, directeur), M. Olivier Soutet (Pr, Université de Paris-Sorbonne, président), M. Patrice Enjalbert (Pr, Université de Caen, rapporteur), Mme Lita Lundquist (Pr Copenhagen Business School, Danemark, rapporteur), M. Laurent Gosselin (Pr, Université de Rouen, examinateur), Mme Tuija Virtanen (Pr, Université Åbo Akademi, Finlande, examinatrice), M. Pierre Zweigenbaum (DR., LIMSI-CNRS, examinateur).

Résumé *: De quoi parle-t-on quand on propose de traiter (de) la temporalité dans les textes ? Dans le paradigme applicatif du traitement automatique des langues, on y associe toujours, semble-t-il, certaines finalités. Dans le paradigme linguistique, on tente de définir la temporalité comme un objet d'analyse. Au cours des travaux que nous avons menés, nous avons perçu tout l'intérêt de chercher à allier ces deux perspectives pour une visée opérationnelle à caractère exploratoire de ce champ. Appréhender la temporalité linguistique en recourant aux catégories du temps, de l'aspect et de la modalité d'une part et à la dimension énonciative d'autre part figure comme une approche « classique » pour la linguistique énonciative en particulier. Les études se concentrent néanmoins sur le niveau phrastique ; il reste alors à préciser les modes opératoires de description de ces quatre dimensions d'analyse à un niveau textuel. C'est la voie de recherche que nous explorons. Elle vise à mieux cerner le principe d'une certaine dynamique dans les mécanismes de référenciation temporelle opérés au sein des textes ; dynamique dont nous présentons ici les premiers éléments de modélisation en proposant de distinguer différents niveaux de référence temporelle et des modes de parcours possibles entre ceux-ci. Concevoir par ailleurs la caractérisation des besoins des utilisateurs finaux comme une part essentielle du processus de modélisation de la temporalité constitue pour nous une gageure à plus long terme, tant théorique qu'applicative. Nous l'envisageons à ce jour* via *le développement d'outils interactifs qui visent à instaurer différentes formes de « lectures temporelles » des textes.*

URL où le mémoire d'HDR pourra être téléchargé :
http : //tel.archives-ouvertes.fr/docs/00/45/24/64/PDF/HDR_DelphineBattistelli_siteHAL.pdf

---

Nabil HATHOUT (nabil.hathout@univ-tlse2.fr)

Titre : Contributions à la description de la structure morphologique du lexique et à l'approche extensive en morphologie.

Mots-clés : morphologie théorique, paradigmes, proximité morphologique, contraintes, analogie, morphologie informatique, préfixation, parasynthèse.

Title : *Contributions to the description of the morphological structure of the lexicon and to extensive morphology.*

Keywords : *theoretical morphology, paradigms, morphological relatedness, constraints, analogy, computational morphology, prefixation, parasynthesis.*

Mémoire de HDR en Linguistique, Université Toulouse2-Le Mirail, Cognition, Langues, Langage, Ergonomie (UMR 5263), Équipe de Recherche en Syntaxe et

Sémantiquesous la direction de Jacques Durand (Pr, Université de Toulouse2). Thèse soutenue le 04/12/2009.

Jury : M. Jacques Durand (Pr, Université Toulouse 2 Le Mirail, directeur), Mme Marie-Paule Péry-Woodley (Pr, Université Toulouse 2, présidente), Mme Dany Amiot (Pr, Université de Lille 3, rapporteur), M. Vito Pirrelli (DR, Istituto di Linguistica Computazionale, CNR, Pisa, Italie, rapporteur), M. Michel Roché (Pr émérite, Université de Toulouse 2 Le Mirail, examinateur), M. Pierre Zweigenbaum (DR., LIMSI-CNRS, examinateur).

Résumé *: Les recherches présentées dans mon mémoire d'habilitation relèvent de la morphologie informatique et descriptive. Elles ont pour finalité première la description de la structure morphologique du lexique et sont centrées sur l'analogie et sur l'acquisition de connaissances morphologiques à partir de lexiques et de dictionnaires. Plusieurs analyseurs morphologiques ont été développés. Le premier, DéCor (dérivations pour les corpus), l'a été dans le cadre du projet MorTAL (analyseur morphologique pour le traitement automatique de la langue). Il exploite les analogies formelles pour construire un réseau dérivationnel dans lequel il recherche les bases des mots dérivés. J'ai ensuite affiné cette méthode en utilisant des informations sémantiques contenues dans des dictionnaires de synonymes. Un deuxième analyseur a ainsi été développé pour acquérir des relations dérivationnelles à partir de quadruplets analogiques particuliers, dérivationnels dans l'une de leurs dimensions et synonymiques dans l'autre. Dans un troisième développement, j'ai proposé un nouveau paradigme d'analyse morphologique permettant de se passer totalement de découpage « morceaulogique » et j'ai redéfini la tâche d'analyse morphologique qui devient une analyse globale du lexique et non plus une analyse de mots isolés. L'analyse consiste à découvrir les différents paradigmes qui structurent le lexique et les analogies qui permettent de les interconnecter puis à caractériser les mots construits par leurs positions dans le maillage défini par ces paradigmes interconnectés.*

Mon deuxième axe de recherche est la morphologie extensive, pratique qui consiste à appuyer les descriptions des phénomènes sur des corpus d'exemples aussi étendus que possible. Ma contribution à cette approche a été multiple : développement de la boîte à outils Webaffix (en collaboration avec Ludovic Tanguy), publication d'articles de synthèse, illustration de la méthode par l'étude de la suffixation en -able *et de la préfixation en* anti-*, et rédaction de* Perl pour les linguistes *(coécrit avec Ludovic Tanguy), un ouvrage destiné aux linguistes qui souhaitent exploiter des données langagières et notamment construire les corpus d'exemples dont ils ont besoin.*

Dans le deuxième chapitre du mémoire, je présente un modèle théorique de la morphologie dérivationnelle. Ce modèle, « lexématique », comporte quatre niveaux de représentation : sémantique, formel, catégoriel et lexical. Le niveau lexical supporte l'organisation morphologique du lexique. L'objectif de la morphologie est de trouver les correspondances les meilleures possibles entre ces quatre niveaux. Ces correspondances sont soumises à un système de contraintes permettant de

sélectionner celles qui offrent la coïncidence optimale entre sens, positions lexicales, formes et catégories. Je présente ensuite huit catégories de paradigmes qui structurent le niveau lexical : les familles et les séries qui peuvent être flexionnelles ou dérivationnelles et morphologiques ou lexicales. Les paradigmes lexicaux sont des extensions des paradigmes morphologiques qui y incluent les supplétions.

J'aborde ensuite dans le troisième chapitre du mémoire les aspects informatiques de mon travail. J'y décris les grandes lignes du nouveau paradigme d'analyse morphologique automatique que je propose. Ce paradigme associe proximité morphologique et analogie formelle pour calculer les relations dérivationnelles. Ce calcul est réalisé sans aucun découpage et sans recourir aux notions de morphème, d'affixe ou d'exposant morphologique. Selon la mesure de proximité morphologique que j'ai définie, deux mots sont d'autant plus proches qu'ils partagent un grand nombre de traits sémantiques et formels et que ces traits sont spécifiques. Cette mesure est calculée en utilisant un algorithme de marche aléatoire dans un bigraphe dont une partie des sommets représente les lexèmes et l'autre leurs propriétés. La mesure de proximité morphologique permet de calculer facilement des voisinages pour un grand nombre de mots, mais elle n'est pas suffisamment fine pour discriminer entre les mots qui sont effectivement apparentés et ceux qui ne le sont pas. Je propose donc de la compléter par la recherche de quadruplets analogiques en exploitant les voisinages morphologiques. Cette seconde technique permet de filtrer finement les voisins mais elle est coûteuse en temps de calcul.

Le quatrième chapitre du mémoire est consacré à la description de la préfixation en anti-*. J'y expose les principales difficultés posées par cette préfixation. La première concerne sa nature catégorielle.* Anti- *est-il un préfixe ou une préposition ? Cette question découle notamment de la grande variété des séquences qui apparaissent derrière* anti-*. L'analyse que je défends est que* anti- *est avant tout un préfixe, même s'il peut marginalement être utilisé comme une préposition. Je présente quelques exemples comme* antitriste*,* anti-obèse *ou* antimordre *qui remettent en cause plusieurs des analyses antérieures de cette préfixation. La seconde question concerne l'existence de plusieurs séries distinctes de dérivés en* anti-*. Je propose de les analyser au moyen de deux critères : (1) l'alternance de l'interprétation endocentrique* vs *exocentrique ; (2) trois modes d'interprétation, spatial, logique et adversatif. Je montre que ces deux critères ne sont pas corrélés et je présente des exemplaires pour cinq des six configurations possibles, notamment des dérivés qui ont un sens spatial exocentrique comme* antisolaire *et des dérivés dont l'interprétation est adversative endocentrique comme* antidésherbant*. La troisième question concerne les dérivés dits parasynthétiques dont je propose une analyse en termes d'emprunt de radicaux. La forme d'un dérivé pourrait en effet être créée en empruntant le radical d'un voisin morphologique lorsque cela permet une meilleure satisfaction de certaines des contraintes qui portent sur la dérivation et notamment de la contrainte de transparence catégorielle.*

URL où le mémoire d'HDR pourra être téléchargé :
http : //w3.erss.univ-tlse2.fr/textes/pagespersos/hathout/publis/Hathout-2009-HDR.pdf

Aurélie PICTON (aurelie.picton@umontreal.ca)

Titre : Diachronie en langue de spécialité. Définition d'une méthode linguistique outillée pour repérer l'évolution des connaissances en corpus. Un exemple appliqué au domaine spatial.

Mots-clés : diachronie, langue de spécialité, linguistique outillée, linguistique de corpus, terminologie textuelle, évolution des connaissances, domaine spatial.

Title : *Diachrony in Terminology. Defining a Tool-based Linguistic Methodology to Trace Knowledge Evolution in Specialized Corpora : A Practical Example from the Field of Space.*

Keywords : *diachrony, languages for special purposes, tool-based linguistics, corpus linguistics, textual terminology, knowledge evolution, field of space.*

Mémoire de thèse de doctorat en Sciences du langage, Université Toulouse 2-Le Mirail, Cognition, Langues, Langage, Ergonomie (UMR 5263), Département de Sciences du Langage/UFR de Langues, Littérature et civilisations étrangères, sous la direction de Anne Condamines (DR, UMR 5263, CLLE-CNRS). Thèse soutenue le 20/10/2009.

Jury : Mme Anne Condamines (DR, UMR CLLE-CNRS, directrice), M. John Humbley (Pr, Université Paris 7, rapporteur), Mme Marie-Claude L'Homme (Pr, Université de Montréal, CA, rapporteur), Mme Nathalie Aussenac-Gilles (CR-HDR, IRIT-CNRS, examinatrice), Mme Pascaline Dury (MC, Université de Lyon 2, examinatrice), M. Daniel Galarreta (Ingénieur, Cnes, examinateur), Mme Marie-Paule Péry-Woodley (Pr, Université de Toulouse 2, examinatrice).

Résumé *: Dans cette thèse, nous menons une réflexion sur la place de la dimension diachronique dans les langues de spécialité, à travers la définition d'une méthode linguistique pour repérer l'évolution des connaissances en corpus.*

Notre recherche s'ancre dans une demande appliquée émanant du Centre national d'études spatiales (CNES), où la question de l'évolution prend une dimension particulière dans le cadre de projets spatiaux dits « de longue durée » (~ 20 ans), au long desquels les connaissances impliquées évoluent nécessairement. Ces projets concernent, par exemple, l'envoi dans l'espace de sondes qui n'atteignent leur destination que plusieurs années après leur lancement. De la même manière, plusieurs générations d'un satellite ou d'un instrument peuvent être développées et nécessitent un suivi de la part des experts. Ces conditions entraînent de nombreuses difficultés dans la pratique des experts : mauvaise communication entre ingénieurs « en poste » et ingénieurs « juniors » qui arrivent en cours de projet, oubli du

contexte des connaissances dans lequel le projet a été initié, plus généralement, modification non consciente du sens et/ou de la forme des termes.
Ce contexte appliqué permet de construire une analyse diachronique en langue de spécialité, perspective largement ignorée en terminologie. Cette situation s'explique par plusieurs raisons, dont les deux principales sont d'ordre théorique et technique. D'un point de vue théorique, la terminologie a en effet longtemps été dominée par la perspective classique wüstérienne, dont l'objectif premier de normalisation exclut la variation temporelle, autrement dit la dimension diachronique. Cela explique que, par conséquent, d'un point de vue technique, très peu de ressources et d'outils existent aujourd'hui pour aborder cette perspective. Néanmoins, depuis les années 1990, la terminologie voit ses fondements théoriques rediscutés et de nouvelles propositions apparaissent qui favorisent la description du fonctionnement du terme, dans toute sa variation. Cette remise en cause est accompagnée et soutenue par les progrès de la linguistique de corpus et du traitement automatique des langues, qui permettent d'explorer de nouvelles voies de recherches jusqu'alors écartées, telles que la diachronie.

Sur la base de ce constat, notre objectif consiste à mettre en place une méthode linguistique et automatisée de repérage de l'évolution des connaissances en corpus, qui réponde à ces besoins appliqués. Pour ce faire, nous prenons appui sur l'expérience de la terminologie textuelle pour l'analyse terminologique. Nous reprenons à notre compte les principes selon lesquels d'une part les connaissances partagées par les experts sont accessibles dans les textes spécialisés et que, d'autre part, les applications liées à la terminologie sont étroitement liées aux textes. Cette position nous permet de poser un parallèle entre langue et connaissances pour observer l'évolution : puisque c'est dans les textes que sont exprimées les connaissances, repérer des évolutions linguistiques dans les textes constitue la clé pour repérer des évolutions de connaissances dans le domaine.

Pour valider cette hypothèse, notre recherche est organisée en trois temps. Tout d'abord, nous identifions et définissons des indices linguistiques repérables en corpus que l'on peut associer de manière régulière à des évolutions de connaissances. Nous proposons quatre indices différents : les empreintes de fréquence, les contextes riches en connaissances évolutives, la coexistence de variantes et les dépendances syntaxiques. La mise au jour de ces indices repose sur une exploration linguistique outillée de deux corpus diachroniques : trois éditions d'un cours d'optique et optoélectronique (1994-2002) et un corpus de rapports de spécification du projet spatial DORIS (1989-2000).
Deuxièmement, nous montrons en quoi chaque indice contribue au repérage d'évolutions potentielles et la manière dont on peut les mettre en œuvre. Plus spécifiquement, nous précisons comment articuler le recours aux outils, les indices eux-mêmes, ainsi que l'interprétation conjointe de l'analyste et de l'expert pour associer évolution dans la langue et évolution des connaissances. La démarche mise en place permet ainsi d'alimenter la question de l'interprétation des données en corpus : tout d'abord à travers la combinaison d'indices comme moyen de construire une interprétation (voire un diagnostic) fiable de l'évolution à partir

d'indices linguistiques, ensuite à travers la question du rôle et de la place des experts dans l'analyse et du dialogue analyste/expert du domaine comme moteur d'une coconstruction de l'interprétation.
Enfin, afin de juger de la pertinence de la méthodologie proposée, nous définissons une typologie des phénomènes d'évolution des connaissances repérés grâce à notre approche. Nous avons pu dégager ainsi dix-sept types d'évolutions différents interprétables en corpus.

Les principales contributions de cette recherche reposent sur l'articulation d'un triple regard pour repérer, étudier et caractériser l'évolution susceptible de se manifester dans les domaines de spécialité, en particulier sur de très courts intervalles temporels. Tout d'abord, d'un point de vue appliqué et méthodologique, notre contribution se situe dans la mise au jour d'une méthode linguistique, automatisée, reproductible et fiable pour repérer l'évolution des connaissances en corpus. Deuxièmement, d'un point de vue descriptif, notre apport se situe dans l'étude de phénomènes d'évolution en jeu en diachronie, développée dans un contexte original dit « en diachronie courte », c'est-à-dire la description de l'évolution sur de très courts intervalles temporels. Enfin, d'un point de vue théorique, notre contribution repose sur la manière d'ancrer l'analyse diachronique en terminologie textuelle et sur l'impact qu'un tel point de vue implique sur le lien entre textes et connaissances et sur la spécificité de l'analyse diachronique dans les langues de spécialité.

URL où la thèse pourra être téléchargée :
http : //tel.archives-ouvertes.fr/tel-00429061/fr/

Anna STAVRIANOU (anna.stavrianou@univ-lyon2.fr)

Titre : Modélisation et fouille de discussions du Web.

Mots-clés : discussions en ligne, fouille de données d'opinion, fouille de texte, réseaux sociaux, systèmes de recommandation, modélisation, forum.

Title : *Modeling and mining of web discussions.*

Keywords : *online discussions, opinion mining, text mining, social networks, recommender systems, modeling, forums.*

Mémoire de thèse de doctorat en Informatique, Université Lumière-Lyon 2, ERIC Laboratoire, Lyonsous la direction de Jean-Hugues Chauchat (Pr, Université Lumière, Lyon) et Julien Velcin (MC, Université Lumière, Lyon). Thèse soutenue le 01/02/2010.

Jury : M. Jean-Hugues Chauchat (Pr, Université Lumière, codirecteur), M. Julien Velcin (MC, Université Lumière, codirecteur), M. Stefan Trausan-Matu (Pr, Université de Bucarest, Roumanie, président), M. Jean-Gabriel Ganascia (Pr, Université de Paris-Sorbonne, rapporteur), M. Pascal Poncelet (Pr, Université de Montpellier, rapporteur), M. Marc El-Bèze (Pr, Université d'Avignon, examinateur).

***Résumé** : Le développement du Web 2.0 a donné lieu à la production d'une grande quantité de discussions en ligne. La fouille et l'extraction de données de qualité de ces discussions en ligne sont importantes dans de nombreux domaines (industrie, marketing) et particulièrement pour toutes les applications de commerce électronique. Les discussions de ce type contiennent des opinions et des croyances de personnes et cela explique l'intérêt de développer des outils d'analyse efficaces pour ces discussions.*

L'objectif de cette thèse est de définir un modèle qui représente les discussions en ligne et facilite leur analyse. Nous proposons un modèle fondé sur des graphes. Les sommets du graphe représentent les objets de type message. Chaque objet de type message contient des informations comme son contenu, son auteur, l'orientation de l'opinion qui y était exprimée et la date où il a été posté. Les liens parmi les objets de type message montrent une relation de type « répondre à ». En d'autres termes, ils montrent quels objets répondent à quoi, conséquence directe de la structure de la discussion en ligne.

Avec ce nouveau modèle, nous proposons un certain nombre de mesures qui guident la fouille au sein de la discussion et permettent d'extraire des informations pertinentes. Les mesures sont définies par la structure de la discussion et la façon dont les objets de type message sont liés entre eux. Il existe des mesures centrées sur l'analyse de l'opinion qui traitent de l'évolution de l'opinion au sein de la discussion. Nous définissons également des mesures centrées sur le temps, qui exploitent la dimension temporelle du modèle, alors que les mesures centrées sur le sujet peuvent être utilisées pour mesurer la présence de sujets dans une discussion.

La représentation d'une discussion en ligne de la manière proposée permet à un utilisateur de « zoomer » dans une discussion. Une liste de messages clés est recommandée à l'utilisateur pour permettre une participation plus efficace au sein de la discussion.

De plus, un système prototype a été implémenté pour permettre à l'utilisateur de fouiller les discussions en ligne en sélectionnant un sous-ensemble d'objets de type message et de naviguer à travers ceux-ci de manière efficace.

URL où la thèse pourra être téléchargée :
http : //recherche.univ-lyon2.fr/eric/sites/eric/IMG/pdf/thesisStavrianou.pdf