Résumés de thèses

Rubrique préparée par Fiammetta Namer

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

Mohamed BELGACEM: (mohamed.belgacem3@voila.fr)

Titre : Reconnaissance automatique de la parole et ALAO : vers un système d'apprentissage de l'arabe oral

Mots-clés : reconnaissance automatique de la parole arabe, apprentissage de langues assisté par ordinateur (ALAO), corpus vocal de l'arabe, modèle acoustique de l'arabe, dictionnaire de prononciation arabe, phonétiseur automatique pour l'arabe, identification automatique des dialectes arabes.

Title: *Automatic speech recognition and CALL : toward a learning system of spoken Arabic*

Keywords: *Automatic Speech Recognition (ASR), Speaker Identification, Computer Assisted Language Learning (CALL).*

Thèse de doctorat en Informatique Linguistique, LIDILEM, Département d'informatique pédagogique, Université Stendhal, Grenoble, sous la direction de Georges Antoniadis (MC, HDR, Université de Grenoble). Thèse soutenue le 16/12/2011.

Jury : M. Georges Antoniadis (MC-HDR, Université Stendhal, Grenoble, directeur), M. Ahmed Jerraya, (DR, CEA-LETI, MINATEC Grenoble France, président), M. Thierry Chanier (Pr, Université Blaise-Pascal, rapporteur), M. Mounir Zrighi (MC-HDR, Université de Monastir, Tunisie, rapporteur), M. Laurent Besacier (Pr, Université Joseph-Fourier, Grenoble, examinateur), Mme Lynne Franje (MC, Université Stendhal, Grenoble, examinatrice).

Résumé : *La composante phonétique n'est pas suffisamment traitée dans l'apprentissage des langues étrangères, pourtant cette dernière est une des sources de la diversité d'accents observés chez les locuteurs non natifs. Le signal de parole nous transmet des informations importantes sur le locuteur. Ces informations nous renseignent sur son niveau de langue et les problèmes de prononciation qu'il peut*

avoir. Une analyse des caractéristiques de la parole des apprenants peut contribuer à améliorer le processus d'apprentissage des langues étrangères. Cette analyse peut aider à la réalisation des applications technologiques comme la reconnaissance automatique de la parole (RAP), la vérification et l'identification du locuteur, l'apprentissage des langues assisté par ordinateur (ALAO). Ces applications ont connu ces dernières années d'importantes avancées grâce à l'évolution des technologies informatiques. Basé sur des ressources issues du traitement automatique des langues (TAL), notre prototype 3AO, élaboré dans cette thèse, se veut d'apporter de nouvelles voies à l'apprentissage de la langue arabe assisté par ordinateur. Un de ses objectifs principaux est de donner aux apprenants de langue arabe des outils adaptés à leurs problématiques didactiques en reléguant au second plan les aspects informatiques.

Dans cette thèse, nous présentons ci-après des outils pour l'apprentissage de la langue arabe assisté par ordinateur. Parmi ces outils il y a la reconnaissance automatique de la parole arabe qui est basée sur la modélisation statistique acoustique ainsi que sur les modèles de Markov cachés (HMM) qui sont aujourd'hui utilisés dans un très grand nombre de systèmes de reconnaissance automatique de la parole. On trouve notamment comme autres outils le corpus vocal, le dictionnaire de prononciation arabe, le phonétiseur pour l'arabe, la technique de l'alignement forcé, l'identification automatique des dialectes arabes, etc.

Nous présentons enfin notre prototype d'apprentissage de l'arabe assisté par ordinateur (3AO), les résultats de nos expériences et la procédure de correction qui permet à l'étudiant de cerner et de corriger ses fautes de prononciation.

URL où la thèse pourra être téléchargée :
http://w3.u-grenoble3.fr/lidilem/labo/web/presentation.php

Ismaïl EL MAAROUF : (elmaarouf.ismail@yahoo.fr)

Titre : Formalisation de connaissances à partir de corpus : modélisation linguistique du contexte pour l'extraction automatique de relations sémantiques

Mots-clés : traitement automatique des langues, linguistique, linguistique de corpus, corpus, sémantique, relation sémantique, extraction d'information, entités nommées, genre textuel, segmentation discursive, désambiguïsation, extraction de patron, acquisition de connaissances à partir de corpus, adaptation de systèmes.

Title: *Corpus-based knowledge formalization: context linguistic modeling for automatic semantic relation extraction*

Keywords: *natural language processing, linguistics, corpus linguistics, corpus, semantics, semantic relations, information extraction, named entities, text genre,*

discourse segmentation, disambiguation, pattern extraction, corpus-driven knowledge acquisition, system adaptation.

Thèse de doctorat en Sciences du Langage, EA 2593 VALORIA, Département MIS, Université de Bretagne Sud, Vannes, sous la direction de Jeanne Villaneau (MC, Université de Bretagne Sud). Thèse soutenue le 06/12/2011.

Jury : Mme Jeanne Villaneau (MC, Université Bretagne Sud, directrice), M. Pierre-François Marteau, (Pr, Université de Bretagne Sud, président), M. Thierry Poibeau (DR, LATTICE-CNRS, rapporteur), Mme Sophie Rosset (CR, LUMSI-CNRS rapporteur), M. Jean-Yves Antoine (Pr, Université de Tours, examinateur).

Résumé : *Les corpus, collections de textes sélectionnés dans un objectif spécifique, occupent une place de plus en plus déterminante en linguistique comme en traitement automatique des langues (TAL). Considérés à la fois comme source de connaissances sur l'usage authentique des langues, ou sur les entités que désignent des expressions linguistiques, ils sont notamment employés pour évaluer la performance d'applications de TAL. Les critères qui prévalent à leur constitution ont un impact évident, mais encore délicat à caractériser, sur (i) les structures linguistiques majeures qu'ils renferment, (ii) les connaissances qui y sont véhiculées, et (iii) la capacité de systèmes informatiques à accomplir une tâche donnée.*

Ce mémoire étudie des méthodologies d'extraction automatique de relations sémantiques dans des corpus de textes écrits. Un tel sujet invite à examiner en détail le contexte dans lequel une expression linguistique s'applique, à identifier les informations qui déterminent son sens, afin d'espérer relier des unités sémantiques. Généralement, la modélisation du contexte est établie à partir de l'analyse de cooccurrences d'informations linguistiques issues de ressources ou obtenues par des systèmes de TAL. Les intérêts et les limites de ces informations sont évalués dans le cadre de la tâche d'extraction de relations sur des corpus de genres différents (article de presse, conte, biographie). Les résultats obtenus permettent d'observer que pour atteindre une représentation sémantique satisfaisante ainsi que pour concevoir des systèmes robustes, ces informations ne suffisent pas.

Deux problèmes sont particulièrement étudiés. D'une part, il semble indispensable d'ajouter des informations qui concernent le genre du texte. Pour caractériser l'impact du genre sur les relations sémantiques, une méthode de classification automatique, reposant sur les restrictions sémantiques qui s'exercent dans le cadre de relations verbo-nominales, est proposée. La méthode est expérimentée sur un corpus de conte et un corpus de presse.

D'autre part, la modélisation du contexte pose des problèmes qui relèvent de la variation discursive de surface. Un texte ne met pas toujours bout à bout des expressions linguistiques en relation et il est parfois nécessaire de recourir à des algorithmes complexes pour détecter des relations à longue portée. Pour répondre à ce problème de façon cohérente, une méthode de segmentation discursive, qui

s'appuie sur des indices de structuration de surface apparaissant dans des corpus écrits, est proposée. Elle ouvre le champ à la conception de grammaires qui permettent de raisonner sur des catégories d'ordre macro-syntaxique afin de structurer la représentation discursive d'une phrase. Cette méthode est appliquée en amont d'une analyse syntaxique et l'amélioration des performances est évaluée. Les solutions proposées à ces deux problèmes nous permettent d'aborder l'extraction d'information sous un angle particulier : le système implémenté est évalué sur une tâche de correction d'entités nommées dans le contexte d'application des systèmes de question-réponse. Ce besoin spécifique entraîne l'alignement de la définition d'une catégorie sur le type de réponse attendue par une question.

URL où la thèse peut-être téléchargée :
http://tel.archives-ouvertes.fr/tel-00657708/fr/

Ludovic JEAN-LOUIS : (ludovic.jeanlouis@gmail.com)

Titre : Approches supervisées et faiblement supervisées pour l'extraction d'événements et le peuplement de bases de connaissances

Mots-clés : extraction d'information, extraction de relations, extraction d'événements.

Title: *Supervised and weakly supervised approaches for event extraction and knowledge base population*

Keywords : *information extraction, relation extraction, event extraction.*

Thèse de doctorat en Informatique, UMR LIMSI-CNRS. UFR Sciences, Université Paris Sud, Orsay. Sous la direction de M. Olivier Ferret (DR, CEA-LIST) et M. Romaric Besançon (CR, CEA-LIST). Soutenue le 15/12/2011.

Jury : M. Olivier Ferret (DR, CEA-LIST, directeur), M. Romaric Besançon (CR, CEA-LIST, co-directeur), M. Patrice Bellot (Pr, Université Aix-Marseille, rapporteur), Mme Adeline Nazarenko (Pr, Université Paris-Nord, rapporteur), M. Claude de Loupy (ingénieur sénior, Syllabs, examinateur), M. Pierre Zweigenbaum (DR, LIMSI-CNRS, examinateur).

Résumé : *La plus grande partie des informations disponibles librement sur le Web se présentent sous une forme textuelle, c'est-à-dire non-structurée. Dans un contexte comme celui de la veille, il est très utile de pouvoir présenter les informations présentes dans les textes sous une forme structurée en se focalisant sur celles jugées pertinentes vis-à-vis du domaine d'intérêt considéré. Néanmoins, lorsque l'on souhaite traiter ces informations de façon systématique, les méthodes*

manuelles ne sont pas envisageables du fait du volume important des données à considérer.
L'extraction d'information s'inscrit dans la perspective de l'automatisation de ce type de tâches en identifiant dans des textes les informations concernant des faits (ou événements) afin de les stocker dans des structures de données préalablement définies. Ces structures, appelées templates *(ou formulaires), agrègent les informations caractéristiques d'un événement ou d'un domaine d'intérêt représentées sous la forme d'entités nommées (nom de lieux, etc.).*

Dans ce contexte, le travail de thèse que nous avons mené s'attache à deux grandes problématiques :
– l'identification des informations liées à un événement lorsque ces informations sont dispersées à une échelle textuelle en présence de plusieurs occurrences d'événements de même type ;
– la réduction de la dépendance vis-à-vis de corpus annotés pour la mise en œuvre d'un système d'extraction d'information.
Concernant la première problématique, nous avons proposé une démarche originale reposant sur deux étapes. La première consiste en une segmentation événementielle identifiant dans un document les zones de texte faisant référence à un même type d'événements, en s'appuyant sur des informations de nature temporelle. Cette segmentation détermine ainsi les zones sur lesquelles le processus d'extraction doit se focaliser.
La seconde étape sélectionne à l'intérieur des segments identifiés comme pertinents les entités associées aux événements. Elle conjugue pour ce faire une extraction de relations entre entités à un niveau local et un processus de fusion global aboutissant à un graphe d'entités. Un processus de désambiguïsation est finalement appliqué à ce graphe pour identifier l'entité occupant un rôle donné vis-à-vis d'un événement lorsque plusieurs sont possibles.
La seconde problématique est abordée dans un contexte de peuplement de bases de connaissances à partir de larges ensembles de documents (plusieurs millions de documents) en considérant un grand nombre (une quarantaine) de types de relations binaires entre entités nommées. Compte tenu de l'effort représenté par l'annotation d'un corpus pour un type de relations donné et du nombre de types de relations considérés, l'objectif est ici de s'affranchir le plus possible du recours à une telle annotation tout en conservant une approche par apprentissage. Cet objectif est réalisé par le biais d'une approche dite de supervision distante prenant comme point de départ des exemples de relations issus d'une base de connaissances et opérant une annotation non supervisée de corpus en fonction de ces relations afin de constituer un ensemble de relations annotées destinées à la construction d'un modèle par apprentissage. Cette approche a été évaluée à large échelle sur les données de la campagne TAC-KBP 2010.

URL où la thèse peut-être téléchargée :
http://tel.archives-ouvertes.fr/tel-00686811

Mathieu MOREY : (mathieu.morey@gmail.com)

Titre : Étiquetage grammatical symbolique et interface syntaxe-sémantique des formalismes grammaticaux lexicalisés polarisés

Mots-clés : étiquetage grammatical, interface syntaxe-sémantique, réécriture de graphes, polarités, grammaires d'interaction, structures de dépendance.

Title: *Symbolic supertagging and syntax-semantics interface of polarized lexicalized grammatical formalisms*

Keywords: *supertagging – syntax-semantics interface – graph rewriting – polarities – interaction grammars – dependency structures*

Thèse de doctorat en Informatique. UMR7503 LORIA Nancy. UFR Mathématiques et informatique, Université Nancy2, sous la direction de Guy Perrier (Pr, Université de Nancy2).

Jury : M. Guy Perrier (Pr, Université de Nancy2, président), M. Alain Polguère (Pr, Université de Nancy2, président), M. Philippe Blache (DR, LPL-CNRS & Université Provence Aix-Marseille1, rapporteur), M. Alexis Nasr (Pr, Université Aix-Marseille2, rapporteur), M. Guillaume Bonfante (MC, INPL, examinateur), M. Gérard Huet (DR, INRIA-Rocquencourt, examinateur), M. Sylvain Kahane (Pr, Université Paris 10 – Nanterre, examinateur).

Résumé : *Les travaux de cette thèse portent sur l'analyse syntaxique et sémantique de la phrase, en utilisant pour l'analyse syntaxique un formalisme grammatical lexicalisé polarisé et en prenant comme exemple les grammaires d'interaction. Dans les formalismes grammaticaux lexicalisés, les polarités permettent de contrôler explicitement la composition des structures syntaxiques. Nous exploitons d'abord le besoin de composition exprimé par certaines polarités pour définir une notion faible de réduction de grammaire applicable à toute grammaire lexicalisée polarisée. Nous étudions ensuite la première phase de l'analyse syntaxique des formalismes lexicalisés : l'étiquetage grammatical. Nous exploitons, là encore, le besoin de composition de certaines polarités pour concevoir trois méthodes symboliques de filtrage des étiquetages grammaticaux que nous implantons sur automate. Nous abordons enfin l'interface syntaxe-sémantique des formalismes lexicalisés. Nous montrons comment l'utilisation de la réécriture de graphes comme modèle de calcul permet concrètement d'utiliser des structures syntaxiques riches pour calculer des représentations sémantiques sous-spécifiées.*

URL où la thèse peut-être téléchargée :
http://tel.archives-ouvertes.fr/tel-00640561

Tantely Harinjaka RAVELONJATOVO : (tantelyh@hotmail.com)

Titre : Contribution a la méthodologie d'analyse systématique des termes malgaches. Cas du domaine de l'environnement

Mots-clés : linguistique de corpus, terminologie textuelle, patrons terminologiques, terminotique, terminologisation, relation syntaxique, relation actancielle, lexicologie.

Title: *Contribution to systematic analysis methodology of malagasy terms. The case of the environnement field*

Keywords: *corpus linguistic, textual terminology, terminological patterns, terminotics, terminologisation, syntactic relation, actancial relation, lexicology.*

Thèse de doctorat en Lettres, option Linguistique appliquée. Centre interdisciplinaire de Recherche Appliquée en Malgache, Département de Langue et Lettres Malgaches, Faculté des Lettres et Sciences Humaines – Université d'Antananarivo, Antananarivo, Madagascar. Sous la direction de Baholisoa Simone Ralalaoherivony (MC, Université d'Antananarivo) et de Béatrice Daille (Pr, Université de Nantes).

Jury : Mme Baholisoa Simone Ralalaoherivony (MC, Université d'Antananarivo, codirectrice), Mme Lucie Rabaovololona Raharinirina (MC, Université d'Antananarivo, présidente), Mme Suzy Rajaonarivo (MC, Université d'Antananarivo, rapporteur), M. Julien Amédée Raboanary (Pr, Institut Supérieur Polytechnique de Madagascar, rapporteur), M. Jeannot Fils Ranaivoson (MC, Université d'Antananarivo, examinateur).

Résumé : *Cette thèse s'inscrit dans le cadre de la recherche appliquée au domaine professionnel de l'environnement. Les trois principaux objectifs sont :*
– l'utilisation effective de la terminologie textuelle pour le malgache ;
– l'indentification des patrons terminologiques malgaches du domaine de l'environnement ;
– l'étude de la formation de ces patrons terminologiques.
Suivant l'hypothèse générale, la création ou la formation des termes malgaches dans le domaine de l'environnement se fait sur la base phrastique. Cette hypothèse se décline en deux séries d'hypothèses spécifiques à savoir les hypothèses par rapport au cadre théorique et les hypothèses par rapport à la formation des termes. Suivant le cadre théorique, la phrase et ses constituants doivent être analysés dans un corpus textuel. La formation des termes s'opère selon des règles syntaxiques

d'une phrase malgache. Ainsi, un terme du domaine de l'environnement est une unité lexicale spécialisée du domaine de l'environnement (ULSE).
La thèse se subdivise en trois parties. La première partie est consacrée à l'état de l'art sur la terminologie textuelle. La deuxième partie est dédiée à l'application du cadre théorique adopté, dans le domaine de l'environnement et en langue malgache. La troisième partie comprend la synthèse sur les acquis de la thèse et quelques exemples d'application possible.
Le domaine de l'environnement malgache se caractérise par la pratique terminologique plutôt que par la théorie. Cela se manifeste par les productions textuelles par les locuteurs du domaine et l'élaboration de certains lexiques y afférents. Les recherches théoriques effectuées en la matière sont peu nombreuses.
Sur le plan international, la terminologie textuelle est un cadre théorique issu de deux écoles de pensée : la linguistique de corpus, notamment des écoles contextualistes britanniques, et la terminologie moderne du cercle de Vienne.
Comme le montre la deuxième partie de la thèse, la mise en application de ce cadre n'a pas été évidente pour une langue peu dotée comme le malgache.
Pour l'analyse effective des termes du domaine, un corpus et un échantillon des termes ont donc été constitués comme matériau d'étude. Un corpus textuel de 500 000 mots, dénommé TONTONA, ainsi qu'un échantillon de 500 termes, ont été construits sur la base de critères linguistiques et extralinguistiques préalablement définis. TONTONA a permis l'indentification des quinze patrons morphosyntaxiques du domaine de l'environnement et deux niveaux de formations terminologiques.
Les quinze patrons identifiés sont répartis en quatre patrons pour les termes simples et onze patrons pour les termes complexes. Ces patrons se forment à la fois au niveau linguistique et au niveau terminologique. Au niveau linguistique, la formation est syntaxique, c'est-à-dire que les termes se forment soit par la relation actanctielle (relation interne des termes), soit par la relation non actanctielle (dérivation syntaxique au sens de Mel'cŭk). Au niveau terminologique, chaque ULSE ainsi créée se terminologise dans les textes du corpus par l'intermédiaire de l'homogénéité conceptuelle de ses cooccurrents, son ancrage au domaine et sa morphologie invariable.
La troisième partie de la thèse récapitule les acquis (théoriques et méthodologiques) et formule un bilan de l'application des patrons terminologiques au domaine de l'agriculture durable.
La plupart des hypothèses avancées ont été vérifiées. Des problèmes ont été tout de même rencontrés lors de la mise en œuvre de la terminologie textuelle. Entre autres, l'exploration de TONTONA a été faite de manière semi-automatique à défaut de logiciel spécifique à la langue malgache. Ainsi, nous ne sommes pas en mesure d'affirmer que tous les termes ont été extraits. Il s'agit d'un échantillon des termes du domaine de l'environnement.

URL où la thèse peut-être téléchargée :
Contacter l'auteur.

Coralie ROTENAUER : (coralie.reutenauer@atilf.fr)

Titre : Vers un traitement automatique de la néosémie : approche textuelle et statistique

Mots-clés : néologie sémantique, nouveau sens, lexique, dictionnaire, textométrie, corpus, acquisition automatique, traits sémantiques, indices statistiques, spécificités, description sémantique multiniveaux.

Title: *Automating meaning acquisition: a textual and statistical approach*

Keywords: *neology, new meaning, meaning aquisition, dictionary, corpus, statistical score, lexical processing, multilevel sense description.*

Thèse de doctorat en Sciences du Langage. ATILF, Université de Lorraine, ED LTS, Nancy. Sous la direction de Jean-Marie Pierrel (Pr, Université de Lorraine), Evelyne Jacquey (CR, ATILF-CNRS), Mathieu Vallette (Pr, INALCO). Soutenue de 20/01/2012.

Jury : M. Jean-Marie Pierrel (Pr, Université de Lorraine, directeur), Mme Evelyne Jacquey (CR, ATILF-CNRS, codirectrice), M. Mathieu Vallette (Pr, INALCO, codirecteur), M. Alain Polguère (Pr, Université de Lorraine, président), Mme Anne Condamine (DR, CLLE-ERSS, rapporteur), M. Jean-François Sablayrolles (Pr, Université Paris 13, rapporteur), M. Ludovic Lebart (DR, Télécom-Paritech, examinateur).

Résumé : *Un défi du traitement automatique du lexique est l'acquisition de nouveaux sens lexicaux. Certaines unités lexicales sont employées dans de nouveaux contextes, comme* toxique *ou* tsunami *en contexte de crise financière, et de ce fait, leur sens se modifie. Dans cette thèse, nous proposons des éléments de modélisation pour affecter de façon semi-automatique de nouveaux contenus sémantiques à une cible lexicale. Ces éléments de modélisation sont définis en trois temps, à travers la description théorique des phénomènes linguistiques en jeu, l'analyse des ressources exploitables et d'outils adaptés pour les appréhender et enfin une mise en œuvre expérimentale.*

Deux phénomènes linguistiques participent à l'émergence d'un nouveau sens : les variations sémantiques et la néologie. Les variations sémantiques d'une unité lexicale résultent de contrastes entre son sens en langue, tel que codé dans des ressources de référence, et son sens en discours, actualisé par des contextes d'emploi. Les variations sémantiques ciblées sont les variations marquées, telles que le sens actualisé présente une rupture avec le sens codé.

Au critère de rupture s'ajoute un critère de diffusion dans le temps : la variation sémantique participe d'un processus et se combine ainsi à de la néologie. Le nouveau sens est recherché pour de la néologie sémantique, définie à la croisée des

variations sémantiques et de la néologie. Un modèle d'allocation de signifié est établi en couplant des indices identifiés comme fondamentaux dans la détection de la néologie à un appareil théorique issu de la sémantique textuelle.
Dans ce cadre théorique, le sens codé est représenté comme un ensemble structuré de sèmes, unités de sens minimales ; le sens en discours est décrit à l'aide d'unités de granularité sémantique décroissante et à travers des phénomènes de récurrences et regroupements de sèmes.
Les ressources et outils utilisés relèvent du champ de la linguistique de corpus et de la lexicométrie. Les ressources sont de deux types, lexicographiques et textuelles. Les ressources lexicographiques constituent le vivier de sens codés. Une plate-forme en extrait des représentations du sens sous forme d'ensembles de sèmes. Les ressources textuelles sont le lieu d'observation des sens actualisés. La caractérisation des sens émergents dépend de leur structure en domaines et dans le temps. En pratique, la ressource lexicographique utilisée est le Trésor de la Langue Française *informatisé et trois corpus journalistiques des années 2000 relevant de différentes thématiques ont servi de ressources textuelles.*
Divers outils mathématiques, notamment statistiques, se prêtent à l'exploitation des grandes bases de données que sont les ressources lexicographiques et textuelles. La structure des ressources peut se concevoir comme un espace mathématique dépendant de paramètres tels que le temps ou les domaines. Le nouveau sens est recherché à travers un jeu de contrastes qui se traduit par un jeu sur des espaces et sous-espaces mathématiques. Pour chaque découpage de l'espace, il est possible d'extraire des unités saillantes à l'aide d'indices statistiques, dont l'interprétation et la validité sont discutées. Différentes techniques permettent ensuite de structurer les unités identifiées comme significatives à travers la série de contrastes.
Dans une perspective applicative, une procédure d'allocation de signifié est proposée. Elle est accompagnée d'expériences illustratives aux différentes étapes. Le déroulement de la procédure est sous-tendu par des caractérisations de niveaux sémantiques de plus en plus précis, allant des domaines aux unités lexicales puis aux sèmes. Des perspectives complémentaires de la procédure sont ensuite proposées. Celles-ci élargissent les traitements proposés à d'autres objets linguistiques et à des formats de représentation des ressources plus complexes.

URL où la thèse peut-être téléchargée :
http://www.atilf.fr/spip.php?article99