Résumés de thèses

Rubrique préparée par Fiammetta Namer

Université Nancy2 de Nancy, UMR « ATILF »

Fiammetta.Namer@univ-nancy2.fr

Jonathan CHEVELU : (jonathan@chevelu.fr)

Titre : Production de paraphrases pour les systèmes vocaux humain-machine

Mots-clés : traitement du langage naturel (informatique), synthèse automatique de la parole, systèmes conversationnels (informatique), optimisation combinatoire.

Title: *Paraphrase generation for human-machine voice interaction systems*

Keywords: *natural language processing; automatic speech generation.*

Thèse de doctorat en Informatique, GREYC – Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen (CNRS : UMR6072) Université de Caen Basse-Normandie, Caen sous la direction de Yves Lepage (Pr., Université de Caen & Université de Waseda (Japon)). Thèse soutenue le 17/03/2011.

Jury: M. Yves Lepage (Pr., Université de Caen & Université de Waseda, directeur), M. Arnaud Lallouet, (Pr, Université de Caen Basse-Normandie, président), M. Marc Dymetman (DR, Xerox Research Center Europe, rapporteur), M. Olivier Boëffard (Pr, ENSSAT, rapporteur), M. Aurélien Max (MC, Université Paris XI, examinateur), M. Thierry Moudenc (CR, Orange Labs, examinateur).

Résumé : *Cette thèse s'intéresse au lien entre ce qui est prononcé par un humain et le système vocal humain-machine qui le prononce. Plutôt que de proposer des systèmes capables de tout vocaliser, nous envisageons le message comme une variable qui peut être modifiée. L'élément primordial d'un message est son sens. Il est donc possible de changer les mots utilisés si cela conserve le sens du message et améliore les systèmes vocaux. Cette modification s'appelle « production de paraphrases ». Dans cette thèse, nous proposons une étude de la production statistique de paraphrases pour les systèmes vocaux humain-machine. Pour ce faire, nous présentons la conception d'un système de référence et d'une plate-forme d'évaluation en ligne. Nous mettons en lumière les différentes limites de l'approche*

classique et nous proposons un autre modèle fondé sur l'application de règles de transformation. Nous montrons qu'il est nécessaire de prendre en compte l'utilisation souhaitée des paraphrases lors de leur production et de leurs évaluations, et pas uniquement le critère de conservation du sens. Enfin, nous proposons et étudions un nouvel algorithme pour produire des paraphrases, fondé sur l'échantillonnage de Monte-Carlo et l'apprentissage par renforcement. Cet algorithme permet de s'affranchir des contraintes habituelles de l'algorithme de Viterbi et donc de proposer librement de nouveaux modèles pour la paraphrase.

URL où la thèse pourra être téléchargée :

<http://tel.archives-ouvertes.fr/tel-00603750/fr/>

Juyeon KANG : (kjuyeon79@yahoo.fr)

Titre : Problèmes morpho-syntaxiques analysés dans un modèle catégoriel étendu : application au coréen et au français avec une réalisation informatique

Mots-clés : grammaires catégorielles, logique combinatoire, analyseur morpho-syntaxique du coréen et du français, double cas, ordre des mots.

Title: *Morpho-syntactic problems analyzed in an extended categorial model: application to Korean and to French with a development of a categorial parser*

Keywords: *categorial grammars, combinatory logic, korean and french parser, double case, word order.*

Thèse de doctorat en Linguistique Informatique, LaLIC (Langues, Logiques, Informatique et Cognition), Université Paris IV Sorbonne, sous la direction de Jean-Pierre Desclès (Pr, Université Paris IV Sorbonne), et de Shihyeon Pak, (Pr, Université Hankuk des Études Étrangères), soutenue le 28/06/2011.

Jury : M. Jean-Pierre Desclès (Pr., Université de Paris IV, directeur), Mme Shihyeon Pak (Pr, Université Hankuk des Études Étrangères, codirectrice), M. Ismail Biskri (Pr, Université du Québec à Trois Rivières, rapporteur), Mme Aeson Yoon (Pr, Université nationale de Pusan, rapporteur), Mme Injoo Choi-Jonin (Pr, Université de Toulouse II, examinatrice), M. Alain Lecomte (Pr, Université de Paris 8, Examineur).

Résumé : *Ce travail de thèse vise à proposer une analyse formelle de phénomènes langagiers tels que le système casuel, le double cas, la flexibilité de l'ordre des mots, la coordination, la subordination et la thématization, dans deux langues structurellement très distinctes : le coréen et le français. Ces problèmes ont été souvent laissés dans l'ombre ou négligés dans les travaux précédents en*

linguistique informatique. L'analyse morpho-syntaxique du coréen que nous proposons vise les aspects à la fois linguistiques, computationnels et logiques du traitement automatique de cette langue.

Le choix théorique s'est porté sur le formalisme de la grammaire catégorielle combinatoire applicative (GCCA), qui met en œuvre les opérateurs de la logique combinatoire et le calcul fonctionnel des types de Church. Le problème à résoudre est le suivant : étant donné les spécificités du coréen (et d'autres langues extrême-orientales), comme la présence de cas, de constructions « à double cas » et un ordre des mots flexibles, cette langue est-elle analysable avec un formalisme catégoriel ? Et suivant quelle stratégie de calcul ? Nous illustrons les réponses à ces questions par un certain nombre d'exemples. Les analyses formelles proposées dans ce travail permettent ensuite d'examiner la pertinence syntaxique de l'hypothèse dite « anti-anti relativiste », en dégagant certains invariants syntaxiques à partir des opérations de prédication, de détermination, de transposition, de quantification et de coordination.

Par la suite, nous démontrons l'automatisme des modèles formels adoptés en développant un analyseur morpho-syntaxique, l'analyseur ACCG (Applicative Combinatory Categorical Grammar). L'implémentation de la GCCA dans l'analyseur ACCG a permis premièrement d'augmenter sa portée grammaticale en traitant les phénomènes complexes des langues naturelles (coordination, subordination, thématization, identification des suffixes porteurs de double cas, flexibilité de l'ordre des mots). Deuxièmement, nous avons pu réduire certaines ambiguïtés – par exemple, la polysémie de certains suffixes nominaux qui relèvent de plusieurs types syntaxiques selon le contexte – par la stratégie de « calcul quasi incrémental de gauche à droite », et par l'introduction de la méthode d'exploration contextuelle. Notre travail permet enfin de montrer que la GCCA n'est pas seulement linguistiquement pertinente mais aussi aisément implémentable. Ceci a été clairement montré dans les résultats d'évaluation de l'analyseur ACCG sur les textes en français et en coréen, ce qui confirme l'aspect multilingue de la méthode.

Notre travail constitue le préalable à la construction de ressources plus larges et construites selon les principes édictés dans notre modèle.

URL où la thèse pourra être téléchargée : demander à l'auteur.

Kais HADDAR : (kais.haddar@yahoo.fr)

Titre : Spécification et implémentation des ressources lexico-syntaxiques normalisées et analyse des phénomènes linguistiques pour la langue arabe

Mots-clés : lexique LMF, transducteurs NooJ, entité nommée arabe, spécification TDL, grammaire HPSG arabisée.

Title: *specification and implementation of normalized lexico-syntactical resources and analysis of linguistic phenomena for the Arabic language*

Keywords: *LMF lexicon, NooJ transducer, arabic named entity, TDL specification, arabic HPSG grammar.*

Mémoire de HDR en Informatique, département d'informatique et de communication de la faculté des Sciences de Sfax, Tunisie, HDR soutenue le 07/07/2011.

Jury : M. Abdelmajid Ben Hamadou (Pr., Institut Supérieur d'Informatique et de Multimédia de Sfax, président), Mme Lamia Hadrich Belguith (MC, Faculté des Sciences Économiques et de Gestion de Sfax, rapporteur), M. Denis Maurel (Pr, Université de Tours, rapporteur), M. Abdelhamid Abdelwahed (Pr., Faculté de Lettres et de Sciences Humaines de Sfax, examinateur), M. Laurent Romary (Pr, Université de Humboldt, Berlin, examinateur).

Résumé : *Le thème principal des travaux présentés est la spécification et l'implémentation de ressources linguistiques normalisées et l'analyse des phénomènes linguistiques pour la langue arabe. L'objectif est d'assurer une analyse syntaxique fiable et robuste pouvant être intégrée dans plusieurs applications de traitement automatique des langues naturelles (TALN), en particulier en langue arabe. En effet, le développement de ces applications comme un traducteur automatique ou un moteur de recherche d'information, nécessite le passage par plusieurs phases (lexicale, syntaxique...).*

La conception des applications de TALN requiert souvent une séparation nette entre les traitements et les ressources linguistiques. En conséquence, l'efficacité de ces applications est devenue de plus en plus dépendante de la qualité des ressources linguistiques informatisées sur lesquelles elle repose. Pour la langue arabe, les ressources de référence sont encore dans un état précaire de développement et la plupart des travaux réalisés donnent des résultats d'une qualité relative et d'une visibilité limitée.

En outre, dès qu'on entreprend de fusionner des ressources ou de les intégrer à des applications, se pose la question de l'interopérabilité. En effet, la représentation interne d'une ressource doit être en adéquation avec celle des applications qui l'utilisent. Des travaux doivent donc être entrepris afin d'élaborer des descriptions et des représentations lexicales standard qui établissent les bases nécessaires assurant cette interopérabilité.

Dans ce contexte, nos recherches ont pour but premier de faciliter la production de ressources linguistiques utilisées dans la phase de l'analyse syntaxique. Plus particulièrement, nos efforts se concentrent sur des stratégies de normalisation pour la création et l'amélioration de ressources linguistiques, et des méthodes de génération automatique des analyseurs lexicaux et syntaxiques (NooJ, LKB).

Ainsi, les recherches s'articulent autour trois axes principaux, décrit chacun dans un chapitre du mémoire : la modélisation et la construction d'un lexique syntaxique

normalisé arabe, la reconnaissance et la traduction des expressions figées arabes – en particulier des entités nommées, désormais EN – et la spécification et l'élaboration d'une grammaire HPSG de l'arabe.

Dans le premier chapitre, nous avons établi un modèle unifié et normalisé du lexique syntaxique arabe en LMF ISO 24613, que nous avons utilisé pour construire un lexique extensible couvrant les principaux niveaux de description linguistique. Ce modèle garantit la représentation des caractéristiques de la langue arabe. La mise en œuvre du modèle est testée avec l'éditeur LEXUS, conforme à LMF, et étendue au moyen d'un vérificateur de contraintes implémenté en Java. Le prototype de projection d'un lexique syntaxique HPSG vers LMF que nous avons implémenté nous a aidés à récupérer des lexiques HPSG existants, à les fusionner et à les intégrer à d'autres lexiques afin d'obtenir une couverture lexicale plus importante. Ce prototype nous a permis d'enrichir le lexique HPSG utilisé dans la plate-forme LKB.

Dans le deuxième chapitre, la démarche proposée de reconnaissance et de traduction des EN arabes nous a conduits à élaborer un modèle de représentation et un système de transducteurs fondés sur une liste de mots déclencheurs. En outre, l'intégration d'un module de translittération dans le processus de la traduction nous a donné un système de traduction plus souple et plus robuste. Les extensions proposées, comme le traitement du pluriel brisé, ont pour but de mieux gérer les EN et d'évoluer vers une détection plus profonde de ceux-ci.

Dans le troisième chapitre, nous modifions la hiérarchie de types HPSG pour la rendre compatible avec la langue arabe. Nous élaborons une typologie des propositions relatives arabes et de la coordination afin d'étendre la couverture de l'analyseur syntaxique. Celui-ci est testé et évalué dans la plate-forme LKB, au moyen d'un corpus de phrases.

Les démarches préconisées tiennent compte de la qualité des ressources linguistiques informatisées en termes d'extensibilité et d'interopérabilité, propriétés garanties par l'utilisation de normes et de formalismes standardisés (norme LMF, consensus des conférences MUC, transducteurs à états finis, HPSG).

URL où le mémoire pourra être téléchargé : demander à l'auteur.

Stanislas OGER (stanislas.oger@univ-avignon.fr)

Titre : Modèles de langage *ad hoc* pour la reconnaissance automatique de la parole

Mots-clés : reconnaissance automatique de la parole, modèle de langage, Internet, mots hors vocabulaire, théorie des possibilités.

Title: Ad hoc language models for automatic speech recognition

Keywords: *automatic speech recognition, language models, World Wide Web, out-of-vocabulary words, possibility theory.*

Thèse de doctorat en Informatique, LIA (Laboratoire Informatique d'Avignon), Université d'Avignon et des Pays de Vaucluse, Avignon, sous la direction de Georges Linarès (Pr, LIA & Université d'Avignon et des Pays de Vaucluse), soutenue le 30/11/2011.

Jury: M. Georges Linarès (Pr, LIA & Université d'Avignon et des Pays de Vaucluse, directeur), M. Mr. Yannick Estève (Pr, LIUM & Université de Le Mans, président), M. Laurent Besacier (Pr, LIG et Université de Grenoble, rapporteur), M. Kamel Smaïli, (Pr, LORIA & Nancy-Université, rapporteur), M. Alexandre Allauzen (MC, LIMSI, & Université Paris 11, examinateur), M. Pascal Nocéra (MC, LIA & Université d'Avignon, examinateur).

Résumé: *Les trois piliers d'un système de reconnaissance automatique de la parole sont le lexique, le modèle de langage et le modèle acoustique. Le lexique fournit l'ensemble des mots qu'il est possible de transcrire, mots associés à leur prononciation. Le modèle acoustique donne une indication sur la manière dont sont réalisées les unités acoustiques, et le modèle de langage apporte la connaissance sur la manière dont les mots s'enchaînent. Dans les systèmes markoviens de reconnaissance automatique de la parole, les modèles acoustiques et linguistiques sont de nature statistique. Leur estimation nécessite de gros volumes de données sélectionnées, normalisées et annotées.*

À l'heure actuelle, les données disponibles sur le Web constituent de loin le plus gros corpus textuel disponible pour les langues française et anglaise. Ces données peuvent potentiellement servir à la construction du lexique et à l'estimation et l'adaptation du modèle de langage. Le travail présenté ici consiste à proposer de nouvelles approches permettant de tirer parti de cette ressource.

Ces travaux s'articulent autour de deux axes. Le premier traite de l'utilisation des données présentes sur le Web pour mettre à jour dynamiquement le lexique du moteur de reconnaissance automatique de la parole. L'approche proposée consiste à augmenter dynamiquement et localement le lexique du moteur de reconnaissance automatique de la parole lorsque des mots inconnus apparaissent dans le flux de paroles. Les nouveaux mots sont extraits du Web grâce à la formulation automatique de requêtes soumises à un moteur de recherche. La phonétisation de ces mots est obtenue grâce à un phonétiseur automatique.

Le second axe consiste à proposer une nouvelle manière de considérer l'information que représente le Web, pour ce faire des éléments de la théorie des possibilités sont utilisés afin de la modéliser. Un modèle de langage possibiliste est alors proposé. Il fournit une estimation de la possibilité d'une séquence de mots à partir de connaissances relatives à l'existence de séquences de mots sur le Web. Un modèle probabiliste Web, reposant sur le compte de documents fournis par un moteur de recherche Web, est également présenté. Plusieurs approches permettant de combiner ces modèles avec des modèles probabilistes classiques estimés sur corpus sont proposées. Les résultats montrent que le fait de combiner les modèles

probabilistes et possibilistes donne de meilleurs résultats que les modèles probabilistes classiques. De plus, les modèles estimés à partir des données Web donnent également de meilleurs résultats que ceux estimés sur corpus.

URL où la thèse pourra être téléchargée :

<http://www.stanislas-oger.fr/fr/publications.html>

Delphine TRIBOUT (dtribout@linguist.jussieu.fr)

Titre : Les conversions de nom à verbe et de verbe à nom en français

Mots-clés : morphologie dérivationnelle, conversion, HPSG.

Title: *Noun-to-verb and verb-to-noun conversions in French*

Keywords: *derivational morphology, conversion, HPSG.*

Thèse de doctorat en Linguistique, LLF (Laboratoire de Linguistique Formelle), Université Paris VII, sous la direction de Bernard Fradin (DR, LLF), soutenue le 26/10/2010.

Jury : M. Bernard Fradin (DR, LLF, directeur), Mme Georgette Dal (PR, STL & Université de Lille 3, présidente & rapporteur), Mme Fiammetta Namer (PR, ATILF & Nancy-Université, rapporteur), Mme Anne Abeillé (PR, LLF&Université de Paris 7, examinatrice), Mme Anna Thornton (PR, Università di L'Aquila, examinatrice), M. Olivier Bonami (MC, LLF & Université Paris-Sorbonne, examinateur).

Résumé : *Menée dans le cadre de la morphologie lexématique, cette thèse est consacrée à deux procédés de formation de lexèmes en français : la conversion de nom à verbe et la conversion de verbe à nom. Nous y présentons à la fois les aspects formels et sémantiques des deux procédés et en proposons une modélisation dans le cadre de HPSG.*

Dans un premier temps, nous discutons les deux propriétés caractéristiques de la conversion, que sont l'appartenance des deux lexèmes à des catégories lexicales différentes et leur identité phonologique. Puis nous circonscrivons les données qui relèvent ou non d'une conversion nom à verbe ou verbe à nom au regard de ces deux propriétés.

Dans un deuxième temps, nous traitons la question de l'orientation de la conversion et nous évaluons les critères qui sont traditionnellement utilisés pour déterminer l'orientation d'une paire nom et verbe en relation de conversion. Nous montrons que ces critères ne sont pas satisfaisants et nous en proposons un nouveau plus fiable, fondé sur l'analyse morphologique des lexèmes. Sur la base de ce critère,

nous cherchons à établir s'il existe d'autres indices de l'orientation de la conversion, avant de conclure que, dans la plupart des cas, l'orientation ne peut être établie sur la base de critères linguistiques.

Enfin, nous étudions les propriétés morphophonologiques et sémantiques des deux conversions, et nous proposons une représentation formelle des allomorphes et des propriétés sémantiques au moyen de HPSG.

URL où la thèse pourra être téléchargée :

<http://www.llf.cnrs.fr/Gens/Tribout/these.php>