Résumés de thèses

Rubrique préparée par Sylvain Pogodalla

INRIA, Villers-lès-Nancy, F-54600, France
Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54500, France
CNRS, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54500, France
sylvain.pogodalla@inria.fr

Farah BENAMARA : benamara@irit.fr

Titre : Calculer le sens : mots, phrases et au-delà

Mots-clés : Compréhension des langues naturelles, sémantique, structure du discours, pragmatique.

Title: *Computing Meaning: Words, Sentences and Beyond*

Keywords: *Natural language understanding, semantics, discourse structure, pragmatics.*

Habilitation à diriger des recherches en informatique, IRIT, UMR 5505, sous la direction de Nicholas Asher (DR, CNRS, IRIT, Toulouse). Habilitation soutenue le 07/12/2016.

Jury : M. Nicholas Asher (DR, CNRS, IRIT, Toulouse, directeur), M. Manfred Stede (Pr, Université de Potsdam, Allemagne, rapporteur), Mme Béatrice Daille (Pr, Université de Nantes, rapporteur), M. Thierry Poibeau (DR, CNRS, LATTICE, Paris, rapporteur), M. Paolo Rosso (Pr, Université Polytechnique de Valence, Espagne, examinateur), Mme Yvette Yannick Mathieu (DR, CNRS, LLF, Paris, examinateur), M. Mohand Boughanem (Pr, Université Paul Sabatier, Toulouse, examinateur).

Résumé : *Making computers understand natural language texts, as people usually do, opens a wide range of fascinating possibilities like machines that can answer natural language questions, machines that can detect ironic statements, or machines that can interpret our feelings and emotions towards a certain topic. In order to achieve these possibilities, computers need to understand the kind of information and inferential*

mechanisms needed to associate meaning to linguistic expressions. I will show that this cannot be achieved without considering the context in which a text is uttered.

An intuitive view is to consider the distinctions between the linguistic information formed by morphological, syntactic, or textual material surrounding to the word, and any other contextual information surrounding the utterance. The interaction between these different sources of contextual information provides a set of challenging issues in the semantics-pragmatics interface and has been the center of my research interests since 10 years. I am specifically interested in studying how the treatment of linguistic phenomena, in particular at the discourse level, can benefit natural language understanding systems, and help such systems advance beyond representations that include only bags of words or bags of sentences. My work has been organized around discourse analysis and processing of evaluative language at different linguistic levels of interpretation:

– From words to sentences: exploring how sentiment is expressed and extracted, at the word, phrase, and sentence levels, and focusing on lexical semantics of evaluative expressions and sentiment composition.

– From sentences to discourse: focusing on the study of the role of discourse structure, including coherence and rhetorical relations, to sentiment analysis and Arabic processing.

– From discourse to pragmatic inferences: focusing on the study of four phenomena for which people make inferences in their everyday use of language: (a) inferences arising when answering natural language questions, (b) inferences made to detect implicit evaluations, (c) inferences that are drawn when searching for the figurative meaning of an utterance, and (d) inferences made to detect the future state of affairs or plans a writer wants to achieve.

The results show that incorporating linguistic insights, discourse information, and other contextual phenomena, in combination with the statistical exploitation of data, can result in an improvement over approaches which take advantage of only one of those perspectives.

URL où le mémoire peut être téléchargé :
https://www.irit.fr/~Farah.Benamara/HDR-PrintVersion.pdf

Jirka MARŠÍK : jiri.marsik89@gmail.com

Titre : Les *effects* et les *handlers* dans le langage naturel

Mots-clés : Sémantique formelle, compositionnalité, effets de bord, monades, grammaires catégorielles abstraites, sémantique dynamique.

Title: *Effects and Handlers in Natural Language*

Keywords: *Formal semantics, compositionality, side effects, monads, abstract categorial grammars, dynamic semantics.*

Thèse de doctorat en informatique, Inria Nancy – Grand Est, LORIA, UMR 7503, Université de Lorraine, sous la direction de Philippe de Groote (DR, Inria Nancy – Grand Est, LORIA, UMR 7503, Villers-lès-Nancy) et Maxime Amblard (MC, Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy). Thèse soutenue le 09/12/2016.

Jury : M. Philippe de Groote (DR, Inria Nancy – Grand Est, LORIA, UMR 7503, Villers-lès-Nancy, codirecteur), M. Maxime Amblard (MC, Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, codirecteur), M. Laurent Vigneron (Pr, Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, président), M. Chris Barker (Pr, New York University, États-Unis, rapporteur), M. Hugo Herbelin (DR, Inria Paris, rapporteur), Mme Myriam Quatrini (MC, Université de la Méditerranée, examinateur), Mme Christina Unger (Dr., Universität Bielefeld, Allemagne, examinateur).

Résumé : *Ces travaux s'intéressent à la modélisation formelle de la sémantique des langues naturelles. Pour cela, nous suivons le principe de compositionnalité qui veut que le sens d'une expression complexe soit une fonction du sens de ses parties. Ces fonctions sont généralement formalisées à l'aide du λ-calcul. Cependant, ce principe est remis en cause par certains usages de la langue, comme les pronoms anaphoriques ou les présuppositions. Ceci oblige à soit abandonner la compositionnalité, soit modifier les structures du sens. Dans le premier cas, le sens n'est alors plus obtenu par un calcul qui correspond à des fonctions mathématiques, mais par un calcul dépendant du contexte, ce qui le rapproche des langages de programmation qui manipulent leur contexte avec des effets de bord. Dans le deuxième cas, lorsque les structures de sens sont ajustées, les nouveaux sens ont tendance à avoir une structure de monade. Ces dernières sont elles-mêmes largement utilisées en programmation fonctionnelle pour coder des effets de bord, que nous retrouvons à nouveau. Par ailleurs, s'il est souvent possible de proposer le traitement d'un unique phénomène, composer plusieurs traitements s'avère être une tâche complexe. Nos travaux proposent d'utiliser les résultats récents autour des langages de programmation pour parvenir à combiner ces modélisations par les effets de bord.*

Pour cela, nous étendons le λ-calcul avec une monade qui implémente les effects *et les* handlers*, une technique récente dans l'étude des effets de bord. Dans la première partie de la thèse, nous démontrons les propriétés fondamentales de ce calcul (préservation de type, confluence et terminaison). Dans la seconde partie, nous montrons comment utiliser le calcul pour le traitement de plusieurs phénomènes linguistiques : deixis, quantification, implicature conventionnelle, anaphore et présupposition. Enfin,*

nous construisons une unique grammaire qui gère ces phénomènes et leurs interactions.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-01417467/

Aleksandre MASKHARASHVILI : alexandermaskharashvili@gmail.com

Titre : Modélisation du discours avec les Grammaires Catégorielles Abstraites

Mots-clés : Grammaires catégorielles abstraites, discours, logique, grammaire, sémantique, syntaxe, TAG.

Title: *Discourse Modeling with Abstract Categorial Grammars*

Keywords: *Abstract categorial grammars, discourse, logic, grammar, semantics, syntax, TAG.*

Thèse de doctorat en informatique, Inria Nancy – Grand Est, LORIA, UMR 7503, Université de Lorraine, sous la direction de Philippe de Groote (DR, Inria Nancy – Grand Est, LORIA, UMR 7503, Villers-lès-Nancy) et Sylvain Pogodalla (CR, Inria Nancy – Grand Est, LORIA, UMR 7503, Villers-lès-Nancy). Thèse soutenue le 01/12/2016.

Jury : M. Philippe de Groote (DR, Inria Nancy – Grand Est, LORIA, UMR 7503, Villers-lès-Nancy, codirecteur), M. Sylvain Pogodalla (CR, Inria Nancy – Grand Est, LORIA, UMR 7503, Villers-lès-Nancy, codirecteur), M. Laurent Prévot (Pr, Université Aix-Marseille, rapporteur), M. Matthew Stone (Pr, Rutgers University, Piscataway, New-Jersey, États-Unis, rapporteur), M. Mathieu Constant (Pr, Université de Lorraine, ATILF, Nancy, président), Mme Laurence Danlos (Pr, Université Paris Diderot, examinateur), Mme Annie Foret (MC, Université de Rennes 1, IRISA, examinateur), M. Christian Retoré (Pr, Université de Montpellier, examinateur).

Résumé : *Ce mémoire de thèse traite de la modélisation du discours dans le cadre des Grammaires Catégorielles Abstraites (Abstract Categorial Grammars, ACG). Les ACG offrent un cadre unifié pour la modélisation de la syntaxe et de la sémantique. Nous nous intéressons en particulier aux formalismes discursifs qui utilisent une approche grammaticale pour rendre compte des régularités des structures discursives. Nous étudions plusieurs formalismes grammaticaux qui s'appuient sur les Grammaires d'Arbres Adjoints (Tree-Adjoining Grammars, TAG) : D-LTAG, G-TAG et D-STAG. Dans notre travail, nous proposons un encodage de G-TAG et un encodage de D-STAG. G-TAG est un formalisme introduit pour la génération de textes en langue naturelle à partir de représentations conceptuelles (sémantiques). D-STAG est un formalisme synchrone pour la modélisation de l'interface syntaxe-sémantique du discours. Il a été introduit pour l'analyse et la construction des structures discursives. L'encodage en ACG de G-TAG et de D-STAG permet d'éclairer le problème des connecteurs discursifs médiaux que les formalismes s'appuyant sur TAG ne traitent*

pas, du moins pas par un mécanisme grammatical. En effet, pour prendre en compte ces connecteurs, D-LTAG, G-TAG et D-STAG utilisent tous une étape extra grammaticale. Notre encodage offre au contraire une approche purement grammaticale de la prise en compte de ces connecteurs discursifs. La méthode que nous proposons est générique et peut servir de solution à tout encodage des connecteurs médiaux de formalismes fondés sur les TAG. Notre encodage de G-TAG et de D-STAG se fait avec des ACG de second ordre. Les grammaires de cette classe sont réversibles. Elles recourent aux mêmes algorithmes polynomiaux pour construire les structures d'analyse, que ce soit à partir de chaînes de caractères ou à partir de formules logiques. Ainsi, ces grammaires peuvent être utilisées aussi bien en analyse qu'en génération. Les problèmes d'analyse et de génération avec les encodages de G-TAG et de D-STAG en ACG sont donc de complexité polynomiale.

URL où le mémoire peut être téléchargé :
https://hal.inria.fr/tel-01412765

Nicolas MAZZIOTTA : nicolas.mazziotta@skynet.be

Titre : Représenter la connaissance en linguistique. Observations sur l'édition de matériaux et sur l'analyse syntaxique

Mots-clés : Syntaxe, sémiotique, diagrammes, édition de texte, dépendance.

Title: *Representing Knowledge in Linguistics. Observations on Editing Linguistic Data and on Syntactic Analysis*

Keywords: *Syntax, semiotics, diagrams, edition, dependency.*

Habilitation à diriger des recherches en sciences du langage, PHILLIA, Université Paris Ouest Nanterre – La Défense, sous la direction de Sylvain Kahane (Pr, Université Paris Ouest Nanterre – La Défense). Habilitation soutenue le 09/12/2016.

Jury : M. Sylvain Kahane (Pr, Université Paris Ouest Nanterre – La Défense, directeur), Mme Annie Bertin (Pr, Université Paris Ouest Nanterre – La Défense, président), M. Alain Polguère (Pr, Université de Lorraine, rapporteur), Mme Lene Schøsler (Pr émérite, Københavns Universitet, Copenhague, Danemark, examinateur), M. Achim Stein (Pr, Universität Stuttgart, Allemagne, rapporteur), M. Pierre Swiggers (Pr, Katholieke Universiteit Leuven et Université de Liège, Belgique, examinateur).

Résumé : *Les connaissances générées par la recherche en linguistique sont souvent représentées dans des éditions de matériaux (éditions de textes ou répertoires structurés) et des diagrammes, considérés ici comme des objets visuels. Au travers de la synthèse réflexive de mes propres pratiques (principalement en philologie numérique et en analyse syntaxique du français), j'examine la manière dont ces représentations « inscrivent » la connaissance et l'ancrent dans un support matériel.*

Les inscriptions de connaissances linguistiques posent trois questions fondamentales. La première est celle de la manière dont nous les lisons. Le premier chapitre du volume est centré sur les conventions de représentation et l'organisation de l'information au sein des éditions de matériaux et des diagrammes syntaxiques. Je montre ainsi que la lecture des éditions de matériaux et la lecture des diagrammes reposent sur des activités cognitives comparables.

Deuxièmement, toute construction d'inscription nécessite de choisir ce qui y sera inscrit. Le deuxième chapitre étudie la question de la sélection de la partie des données et de l'analyse qui est représentée dans les inscriptions. Une opération de réduction est toujours nécessaire et accompagne le choix de ce qui est explicité. Ce qui est choisi pour être explicité peut ainsi être matérialisé, devenir un objet dans la représentation (on parlera de « réification »). Telle lettre, telle ligne, tel symbole dont on peut percevoir les contours, représente ainsi un contenu notionnel, mais gagne en outre automatiquement des propriétés visuelles supplémentaires. La question de la sélection porte également sur la formalisation de l'encodage de la connaissance, car les structures choisies pour ce faire contraignent ce qui peut être représenté et, partant, ce qui peut être réifié.

Troisièmement, comment les inscriptions permettent-elles de découvrir davantage, c'est-à-dire de créer une connaissance nouvelle qui dépasse ce qu'elles représentent initialement ? Se focalisant sur l'utilisation des diagrammes, le troisième chapitre montre que ces derniers sont primordiaux pour construire de nouvelles connaissances, tant à propos des données que de la théorie. Les diagrammes déployés dans l'espace autorisent toutes sortes de manipulations graphiques qui permettent de faire évoluer la manière dont on conçoit les données et dont on les classe. Ce potentiel ouvre notamment la porte à l'enrichissement des formalismes graphiques et à la redécouverte de formalismes graphiques plus riches et puissants que les arbres et les graphes : les « polygraphes » (structures comportant des nœuds et des arêtes qui peuvent avoir d'autres arêtes comme sommets).

Suite à ces trois chapitres, un quatrième, transversal, est dédié aux inscriptions numériques, qui renouvellent les réponses qu'on peut apporter à toutes ces questions.

URL où le mémoire peut être téléchargé :
http://hdl.handle.net/2268/204408

Andon TCHECHMEDJIEV : andon.tchechmedjiev@lirmm.fr

Titre : Interopérabilité sémantique multilingue des ressources lexicales en données lexicales liées ouvertes

Mots-clés : Alignement automatique de sens, acceptions interlignes, données lexicales liées ouvertes, DBNary, Ontolex.

Title: *Semantic Interoperability of Multilingual Lexical Resources as Lexical Linked Open Data*

Keywords: *Word sense alignement, interlingual acceptions, lexical linked open data, DBNary, Ontolex.*

Thèse de doctorat en informatique, Laboratoire d'Informatique de Grenoble, école doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique (MSTII), Université Grenoble Alpes, sous la direction de Gilles Sérasset (MC, Université Grenoble Alpes) et Jérôme Goulian (MC, Université Grenoble Alpes). Thèse soutenue le 14/10/2016.

Jury : M. Gilles Sérasset (MC, Université Grenoble Alpes, codirecteur), M. Jérôme Goulian (MC, Université Grenoble Alpes, codirecteur), M. Éric Gaussier (Pr, Université Grenoble Alpes, président), M. Mathieu Lafourcade (MC, Université de Montpellier, rapporteur), M. Roberto Navigli (associate professor, Sapienza Università di Roma, Italie, rapporteur), M. Nabil Hathout (DR, CNRS, CLLE/ERSS, Toulouse, examinateur), M. Denis Maurel (Pr, Université François Rabelais de Tours, examinateur), M. Didier Schwab (MC, Université Grenoble Alpes, invité).

Résumé : *Lorsqu'il s'agit de la construction de ressources lexico-sémantiques multilingues, la première chose qui vient à l'esprit est la nécessité que les ressources à aligner partagent le même format de données et la même représentation (interopérabilité représentationnelle). Avec l'apparition de standards tels que LMF et leur adaptation au web sémantique pour la production de ressources lexico-sémantiques multilingues en tant que données lexicales liées ouvertes (Ontolex), l'interopérabilité représentationnelle n'est plus un verrou majeur, le web sémantique offrant des formats de représentation, des technologies de stockage et d'accès à l'information standard et robustes. Cependant, en ce qui concerne l'interopérabilité sémantique des alignements multilingues, le choix et la construction automatique d'un pivot interlingue est l'un des obstacles principaux.*

Pour nombre de ressources (par exemple BabelNet, EuroWordNet), le choix est fait d'utiliser les sens en langue anglaise, ou une autre langue naturelle, comme pivot interlingue. Ce choix mène à une perte de contraste dans les cas où des sens (dans la même acception) du pivot ont des lexicalisations différentes dans plusieurs autres langues. L'utilisation d'un pivot basé sur des acceptions interlingues, solution proposée il y a déjà plus de 20 ans, pourrait être viable, malgré plusieurs tentatives infructueuses. Néanmoins, leur construction manuelle est très ardue et coûteuse et leur construction automatique pose problème du fait de l'absence d'une formalisation et d'une caractérisation axiomatique permettant de garantir leurs propriétés.

Nous proposons dans cette thèse de d'abord formaliser l'architecture à pivots interlingues par acceptions, en développant une axiomatisation garantissant leurs propriétés et leur construction automatique correcte. Nous proposons ensuite des algorithmes de construction initiale automatique ainsi que de mise à jour, en utilisant les propriétés combinatoires et topologiques du graphe des alignements de sens bilingues (paire

à paire). Dans un deuxième temps, nous étudions les implications de l'application de ces algorithmes sur DBNary, une ressource en données lexicales liées ouvertes extraite à partir de Wiktionary de manière régulière.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-01425123/

Izabella THOMAS : izabella.thomas@univ-fcomte.fr

Titre : Écrire en langues spécialisées : méthodes et outils du traitement automatique des langues au service de l'autonomie des rédacteurs

Mots-clés : Langue spécialisée, langue de spécialité, langue contrôlée, rédaction technique, aide à la rédaction, lexique spécialisé, terminologie, extraction de termes, écrits scientifiques, apprentissage des langues spécialisées, enseignement des langues spécialisées.

Title: *Writing in Specialized Languages: Methods and Tools of Natural Language Processing to Enhance the Autonomy of Technical Writers*

Keywords: *Specialized language, language for specific purposes, LSP, controlled language, technical writing, writing aid, specialized vocabulary, terminology, term extraction, scientific writing, LSP learning, LSP teaching.*

Habilitation à diriger des recherches en sciences du langage et TAL, UFR Sciences du Langage, de l'Homme et de la Société, Université de Franche-Comté, Besançon. Habilitation soutenue le 07/12/2016.

Jury : M. Bohdan Krzysztof Bogacki (Pr, Université de Varsovie, Pologne, président), Mme Pierrette Bouillon (Pr, Université de Genève, Suisse, examinateur), Mme Sylviane Cardey-Greenfield (Pr, Université de Franche-Comté, Besançon, rapporteur), Mme Natalie Kübler (Pr, Université Paris Diderot, rapporteur), M. Christophe Roche (Pr, Université de Savoie, rapporteur), Mme Dominique Angèle Vuitton (Pr, Université de Franche-Comté, examinateur).

Résumé : *Ce mémoire est consacré à la présentation de divers travaux entrepris dans l'objectif de concevoir des outils d'aide à la rédaction de textes en langues spécialisées. Il y est question de l'importance de la modélisation linguistique, des contraintes liées à l'automatisation de certaines tâches, du rôle de l'expert et de la prise en compte des utilisateurs finaux. Le rôle prépondérant est donné au lexique et à sa contextualisation, une thématique, qui, je pense, fait consensus parmi les chercheurs travaillant sur la rédaction en langues spécialisées. J'y présente quatre logiciels qui sont conçus pour donner à un auteur d'un texte spécialisé plus d'autonomie dans la rédaction. Certains de ces logiciels s'adressent directement aux rédacteurs (Compagnon LiSe et SARS, Système d'Aide à la Rédaction Scientifique). D'autres constituent des aides indirectes, soit à l'enseignement et à l'apprentissage du vocabulaire spécialisé du niveau académique, soit à la constitution des lexiques pour la rédaction en langues*

contrôlées (Station Sensunique). Ces logiciels tentent modestement de répondre à un véritable besoin sociétal, à savoir la nécessité de produire des textes techniques de qualité, pour des rédacteurs qui ne sont pas toujours des professionnels de la rédaction. La spécificité de ces rédacteurs occasionnels — qu'il s'agisse de rédaction technique ou scientifique — impose de véritables contraintes sur la conception des outils d'aide à la rédaction, résultant d'injonctions apparemment contradictoires : d'une part, le besoin d'outils simples, mais précis, et d'autre part,la complexité de l'information à transmettre.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-01464298