Résumés de thèses

Rubrique préparée par Sylvain Pogodalla

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sylvain.pogodalla@inria.fr

Marianne DJEMAA : marianne.djemaa@gmail.com

Titre : Stratégie domaine par domaine pour la création d'un FrameNet du français : annotations en corpus de cadres et rôles sémantiques

Mots-clés : Traitement automatique de la langue (TAL), linguistique, FrameNet, French FrameNet, sémantique lexicale, annotation sémantique, rôles sémantiques, corpus, français.

Title: *Domain by Domain Strategy for Creating a French FrameNet: Corpus Annotations of Semantic Frames and Roles*

Keywords: *Natural language processing (NLP), linguistics, FrameNet, French FrameNet, lexical semantics, semantic annotation, semantic roles, corpus, French.*

Thèse de doctorat en Linguistique théorique, descriptive et automatique, École doctorale de Sciences du Langage, Laboratoire de Linguistique Formelle (CNRS et Université Paris Diderot – Paris 7), Université Sorbonne Paris Cité, sous la direction de Marie (MC, Université Paris Diderot – Paris 7). Thèse soutenue le 14/06/2017.

Jury : Mme Marie (MC, Université Paris Diderot – Paris 7, directrice), Mme Laurence Danlos (Pr, Université Paris Diderot – Paris 7, présidente), M. Sylvain Kahane (Pr, Université Paris Ouest – Paris 10, rapporteur), Mme Marie-Claude L'Homme (Pr, Université de Montréal, Canada, rapporteur), M. Alexis Nasr (Pr, Aix-Marseille Université, examinateur).

Résumé : *Dans cette thèse, nous décrivons la création du French FrameNet (FFN), une ressource de type FrameNet pour le français créée à partir du FrameNet de l'anglais et de deux corpus arborés : le French Treebank et le Sequoia Treebank. La ressource séminale, le FrameNet de l'anglais, constitue un modèle d'annotation sémantique de situations prototypiques et de leurs participants. Elle propose à la fois :*

a) un ensemble structuré de situations prototypiques, appelées cadres, associées à des caractérisations sémantiques des participants impliqués (les rôles) ;
b) un lexique de déclencheurs, les lexèmes évoquant ces cadres ;
c) un ensemble d'annotations en cadres pour l'anglais.

Pour créer le FFN, nous avons suivi une approche « par domaine notionnel » : nous avons défini quatre « domaines » centrés chacun autour d'une notion (cause, communication langagière, position cognitive ou transaction commerciale), que nous avons travaillé à couvrir exhaustivement à la fois pour la définition des cadres sémantiques, la définition du lexique, et l'annotation en corpus. Cette stratégie permet de garantir une plus grande cohérence dans la structuration en cadres sémantiques, tout en abordant la polysémie au sein d'un domaine et entre les domaines. De plus, nous avons annoté les cadres de nos domaines sur du texte continu, sans sélection d'occurrences : nous préservons ainsi la distribution des caractéristiques lexicales et syntaxiques de l'évocation des cadres dans notre corpus. À l'heure actuelle, le FFN comporte 105 cadres et 873 déclencheurs distincts, qui donnent lieu à 1109 paires déclencheur-cadre distinctes, c'est-à-dire 1109 sens. Le corpus annoté compte au total 16167 annotations de cadres de nos domaines et de leurs rôles.

La thèse commence par resituer le modèle FrameNet dans un contexte théorique plus large. Nous justifions ensuite le choix de nous appuyer sur cette ressource et motivons notre méthodologie en domaines notionnels. Nous explicitons pour le FFN certaines notions définies pour le FrameNet de l'anglais que nous avons jugées trop floues pour être appliquées de manière cohérente. Nous introduisons en particulier des critères plus directement syntaxiques pour la définition du périmètre lexical d'un cadre, ainsi que pour la distinction entre rôles noyaux et non-noyaux.

Nous décrivons ensuite la création du FFN : d'abord, la délimitation de la structure de cadres utilisée pour le FFN, et la création de leur lexique. Nous présentons alors de manière approfondie le domaine notionnel des positions cognitives, qui englobe les cadres portant sur le degré de certitude d'un être doué de conscience sur une proposition. Puis, nous présentons notre méthodologie d'annotation du corpus en cadres et en rôles. À cette occasion, nous passons en revue certains phénomènes linguistiques qu'il nous a fallu traiter pour obtenir une annotation cohérente ; c'est par exemple le cas des constructions à attribut de l'objet.

Enfin, nous présentons des données quantitatives sur le FFN tel qu'il est à ce jour et sur son évaluation. Nous terminons sur des perspectives de travaux d'amélioration et d'exploitation de la ressource créée.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-01661689

Yoann DUPONT : yoa.dupont@gmail.com

Titre : La structuration dans les entités nommées

Mots-clés : Reconnaissance des entités nommées, entités nommées structurées, apprentissage automatique, champs aléatoires conditionnels, réseaux de neurones.

Title: *Structuration in Named Entities*

Keywords: *Named entity recognition, structured named entities, machine learning, conditional random fields, neural networks.*

Thèse de doctorat en Sciences du Langage, Lattice, UMR 8094, Université Sorbonne Nouvelle – Paris 3, sous la direction de Isabelle Tellier (Pr, Université Sorbonne Nouvelle – Paris 3), Christian Lautier (directeur technique, Expert System France), Marco Dinarelli (CR, CNRS). Thèse soutenue le 23/11/2017.

Jury : Mme Isabelle Tellier (Pr, Université Sorbonne Nouvelle – Paris 3, codirectrice), M. Christian Lautier (directeur technique, Expert System France, codirecteur), M. Marco Dinarelli (CR, CNRS, codirecteur), Mme Agata Savary (MC HDR, Université François Rabelais Tours, rapporteur), M. François Yvon (Pr, Université Paris Sud, LIMSI/CNRS, rapporteur), M. Frédéric Landragin (DR, CNRS, examinateur), Mme Pascale Sébillot (Pr, IRISA/INSA de Rennes, examinatrice), M. Patrick Watrin (logisticien de recherche, Université catholique de Louvain, Belgique, examinateur).

Résumé : *La reconnaissance des entités nommées est une discipline cruciale du domaine du TAL. Elle sert à l'extraction de relations entre entités nommées, ce qui permet la construction d'une base de connaissances, le résumé automatique, etc. Nous nous intéressons ici aux phénomènes de structurations qui les entourent.*

Nous distinguons tout d'abord deux types d'éléments structurels dans une entité nommée. Les premiers sont des sous-chaînes récurrentes, que nous appellerons les affixes caractéristiques d'une entité nommée. Le second type d'éléments sont les tokens *ayant un fort pouvoir discriminant, appelés des* tokens *déclencheurs. Nous détaillerons l'algorithme que nous avons mis en place pour extraire les affixes caractéristiques, que nous comparerons à Morfessor. Nous appliquerons ensuite notre méthode pour extraire les* tokens *déclencheurs, utilisés pour l'extraction d'entités nommées du français et d'adresses postales.*

Une autre forme de structuration pour les entités nommées est de nature syntaxique, d'imbrications ou arborée. Pour identifier automatiquement cette structuration, nous proposons un type de cascade d'étiqueteurs linéaires qui n'avait jusqu'à présent jamais été utilisé pour la reconnaissance d'entités nommées. Elles généralisent les approches précédentes qui sont capables de reconnaître uniquement des entités de profondeur limitée ou qui ne peuvent pas modéliser certaines particularités des entités nommées structurées.

Tout au long de cette thèse, nous comparons deux méthodes par apprentissage automatique, à savoir les CRF et les réseaux de neurones, dont nous présenterons les avantages et inconvénients.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-01772268

Mouna ELASTHER : elashtermouna@yahoo.com

Titre : Gestion et extension automatiques du dictionnaire relationnel multilingue de noms propres Prolexbase : Mise à jour multilingue et création d'un volume arabe via la Wikipédia

Mots-clés : Nom propre, Prolexbase, bases lexicales multilingues, notoriété, langue arabe, Wikipédia.

Title: *Automatic Management and Extension of the Multilingual Relational Dictionary of Proper Names Prolexbase: Multilingual Update and Creation of an Arabic Volume via the Wikipedia*

Keywords: *Proper noun, Prolexbase, multilingual lexical databases, notoriety, Arabic language, Wikipedia.*

Thèse de doctorat en Informatique, École doctorale Mathématiques, Informatique, Physique Théorique et Ingénierie des Systèmes, Laboratoire d'Informatique, Université François Rabelais de Tours, sous la direction de Denis Maurel (Pr, Université François Rabelais de Tours). Thèse soutenue le 04/07/2017.

Jury : Mme Béatrice Daille (Pr, Université de Nantes, présidenteet rapporteur), M. Kais Haddar (MC HDR, Université de Sfax, Tunisie, rapporteur), Mme Béatrice Markhoff (MC HDR, Université François Rabelais de Tours, examinatrice), M. Denis Maurel (Pr, Université François Rabelais de Tours, directeur).

Résumé : *Les bases de données lexicales jouent un rôle important dans plusieurs domaines du traitement automatique des langues (TAL), comme l'extraction d'information, la reconnaissance d'entités nommées et la traduction automatique des noms propres. Toutefois, elles nécessitent un développement et un enrichissement permanents par l'exploitation des ressources libres et riches en textes du web sémantique, entre autres, l'encyclopédie universelle Wikipédia, DBpedia, Geonames et Yago2.*

Le dictionnaire électronique relationnel multilingue de noms propres, Prolexbase, issu de nombreux travaux de recherche sur le TAL, comporte à ce jour dix langues, parmi lesquelles trois sont bien couvertes : le français, l'anglais et le polonais. Il a été conçu manuellement et une première tentative semi-automatique a été réalisée par le projet ProlexFeeder. Notre travail avait pour objectif d'élaborer un outil de mise à jour et d'extension automatique de ce lexique, et l'ajout de la langue arabe. Tout d'abord, une mise à jour multilingue de la base de données a été effectuée grâce à l'établisse-

ment d'un système automatique de consolidation des liens Wikipédia dans Prolexbase en nous servant du concept interlangue de Wikipédia. En conséquence, un nombre considérable de nouveaux liens Wikipédia a été ajouté dans toutes les langues constituant la base de données, et cet ajout a été précédé, le cas échéant, d'un traitement des redirections.

Un système entièrement automatique a également été mis en place qui permet de calculer, via l'encyclopédie Wikipédia, un indice de notoriété pour les entrées de Prolexbase. Cet indice dépend de la langue et participe, d'une part, à la construction d'un module de Prolexbase pour la langue arabe et, d'autre part, à la révision de la notoriété actuellement présente pour les autres langues de la base. Pour calculer la notoriété, une technique multicritères de l'aide à la décision a été utilisée : la méthode SAW incluant le calcul de l'entropie de Shannon, à partir de cinq valeurs numériques déduites de l'encyclopédie Wikipédia. Finalement, l'utilisation des liens Wikipédia a été l'instrument fondamental pour la création d'un volume arabe dans Prolexbase par un processus d'extraction de noms propres arabes depuis leurs liens Wikipédia obtenus précédemment.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-01657366

Jérémy FERRERO : ferrero.jerem@gmail.com

Titre : Similarités textuelles sémantiques translingues : vers la détection automatique du plagiat par traduction

Mots-clés : Plagiat, détection de plagiat, détection de similarité, translingue, traduction.

Title: *Cross Lingual Semantic Textual Similarity Detection: towards Cross-Language Plagiarism Detection*

Keywords: *Cross-language, cross-lingual, plagiarism, plagiarism detection, similarity detection.*

Thèse de doctorat en Informatique, Laboratoire Informatique de Grenoble (LIG), UFR Informatique, mathématiques, mathématiques appliquées de Grenoble, Université Grenoble Alpes, sous la direction de Laurent Besacier (Pr, Université Grenoble Alpes). Thèse soutenue le 08/12/2017.

Jury : M. Laurent Besacier (Pr, Université Grenoble Alpes, directeur), M. Didier Schwab (MC, Université Grenoble Alpes, examinateur), Mme Isabelle Tellier (Pr, Université Sorbonne Nouvelle – Paris 3, présidente), M. Emmanuel Morin (Pr, Université de Nantes, rapporteur), M. Juan-Manuel Torres-Moreno (MCHDR, Université d'Avignon et des Pays de Vaucluse, rapporteur), M. Frédéric Agnès (ingénieur, Compilatio, examinateur).

Résumé : *La mise à disposition massive de documents via Internet (pages Web, entrepôts de données, documents numériques, numérisés ou retranscrits, etc.) rend de plus en plus aisée la récupération d'idées. Malheureusement, ce phénomène s'accompagne d'une augmentation des cas de plagiat.*

En effet, s'approprier du contenu, peu importe sa forme, sans le consentement de son auteur (ou de ses ayants droit) et sans citer ses sources, dans le but de le présenter comme sa propre œuvre ou création, est considéré comme plagiat. De plus, ces dernières années, l'expansion d'Internet a également facilité l'accès à des documents du monde entier (écrits dans des langues étrangères) et à des outils de traduction automatique de plus en plus performants, accentuant ainsi la progression d'un nouveau type de plagiat : le plagiat translingue. Ce plagiat implique l'emprunt d'un texte tout en le traduisant (manuellement ou automatiquement) de sa langue originale vers la langue du document dans lequel le plagiaire veut l'inclure. De nos jours, la prévention du plagiat commence à porter ses fruits, grâce notamment à des logiciels antiplagiat performants qui reposent sur des techniques de comparaison monolingue déjà bien éprouvées. Néanmoins, ces derniers ne traitent pas encore de manière efficace les cas translingues. Cette thèse est née du besoin de Compilatio, une société d'édition de l'un de ces logiciels antiplagiat, de mesurer des similarités textuelles sémantiques translingues (sous-tâche de la détection du plagiat).

*Après avoir défini le plagiat et les différents concepts abordés au cours de cette thèse, nous établissons un état de l'art des différentes approches de détection du plagiat translingue. Nous présentons également les différents corpus déjà existants pour la détection du plagiat translingue et exposons les limites qu'ils peuvent rencontrer lors d'une évaluation de méthodes de détection du plagiat translingue. Nous présentons ensuite le corpus que nous avons constitué et qui ne possède pas la plupart des limites rencontrées par les différents corpus déjà existants. Nous menons, à l'aide de ce nouveau corpus, une évaluation de plusieurs méthodes de l'état de l'art et découvrons que ces dernières se comportent différemment en fonction de certaines caractéristiques des textes sur lesquelles elles opèrent. Ensuite, nous présentons de nouvelles méthodes de mesure de similarités textuelles sémantiques translingues basées sur des représentations continues de mots (*word embeddings*). Nous proposons également une notion de pondération morphosyntaxique et fréquentielle de mots, qui peut aussi bien être utilisée au sein d'un vecteur qu'au sein d'un sac de mots, et nous montrons que son introduction dans ces nouvelles méthodes augmente leurs performances respectives. Nous testons ensuite différents systèmes de fusion et combinaison entre différentes méthodes et étudions les performances, sur notre corpus, de ces méthodes et fusions en les comparant à celles des méthodes de l'état de l'art. Nous obtenons ainsi de meilleurs résultats que l'état de l'art dans la totalité des sous-corpus étudiés. Nous terminons en présentant et discutant les résultats de ces méthodes lors de notre participation à la tâche de similarité textuelle sémantique (STS) translingue de*

la campagne d'évaluation SemEval 2017, pour laquelle nous nous sommes classés 1ers pour la sous-tâche correspondant le plus au scénario industriel de Compilatio.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-01721390

Dhaou GHOUL : dhaou.ghoul@gmail.com

Titre : Classifications et grammaires des invariants lexicaux arabes en prévision d'un traitement informatique de cette langue. Construction d'un modèle théorique de l'arabe : la grammaire des invariants lexicaux temporels

Mots-clés : Corpus, classification, désambiguïsation, environnement syntaxique, expression régulière, langue arabe, invariants lexicaux, identification, règles linguistiques, grammaire régulière, schémas de grammaires, TAL.

Title: *Classifications and Grammars of Arab Lexical Invariants in Anticipation of an Automatic Processing of this Language. Construction of a theoretical Model of the Arabic Language: The Temporal Invariants*

Keywords: *Corpus, classification, disambiguation, syntactic environment, regular expression, Arabic language, lexical invariants, identification, linguistic rules, regular grammar, diagrams of grammars, NLP.*

Thèse de doctorat en linguistique, section informatique, Université Paris-Sorbonne, sous la direction de Amr Helmy Ibrahim (Pr émérite, Université Paris-Sorbonne et Université de Franche-Comté, Besançon). Thèse soutenue le 07/12/2016.

Jury : M. Amr Helmy Ibrahim (Pr émérite, Université Paris-Sorbonne et Université de Franche-Comté, Besançon, directeur), M. Mounir Zrigui (Pr, Université de Monastir, Tunisie, rapporteur et président), M. Mohamed Embarki (Pr, Université de Franche-Comté, Besançon, rapporteur), M. André Jaccarini (CR, CNRS, IRAA, examinateur).

Résumé : *Cette thèse porte sur la classification et le traitement des invariants lexicaux arabes qui sont des marqueurs de temporalité et d'aspect. Nous avons associé à chaque invariant des schémas de grammaire (sous forme d'automates). Dans ce travail, nous avons limité notre traitement à vingt invariants lexicaux. Notre hypothèse est construite à partir du principe selon lequel les invariants lexicaux sont situés au même niveau structural que les schèmes dans le langage quotient (squelette) de la langue arabe. Ils recèlent beaucoup d'informations et entraînent des attentes syntaxiques qui permettent de prédire la structure de la phrase. Au début de cette thèse, nous abordons la notion d'*invariant lexical *en exposant les différents niveaux d'invariance. Ensuite, nous classons les invariants étudiés dans cette thèse selon plusieurs critères. La deuxième partie de cette thèse concerne les invariants lexicaux temporels. Nous commençons par une présentation de notre méthode d'étude linguistique ainsi*

que la modélisation par schémas de grammaires associés aux invariants lexicaux temporels étudiés. Ensuite, nous abordons l'analyse proprement dite des invariants lexicaux simples (comme ḥattā*,* baʿda*) et complexes (comme* baʿdamā*,* baynamā*). Enfin, une application expérimentale, Kawâkib, a été employée pour détecter et identifier les contextes de ces invariants lexicaux. Nous montrons les points forts de ses fonctionnalités ainsi que ses lacunes. Nous proposons également une nouvelle vision de la prochaine version de Kawâkib qui peut aussi représenter une application pédagogique de l'arabe avec un recours au lexique minimal.*

URL où le mémoire peut être téléchargé :
http://www.mmsh.univ-aix.fr/program/Documents/GHOUL_Dhaou_2016_These.pdf

---

Pierre-Antoine JEAN : pierreantoine.jean@gmail.com

Titre : Gestion de l'imprécision et de l'incertitude dans un processus d'extraction de connaissances

Mots-clés : Extraction de connaissances, TAL, inférence, représentation des connaissances, incertitude.

Title: *Imprecision and Uncertainty Management in a Knowledge Extraction Process*

Keywords: *Knowledge extraction, NLP, inference, knowledge representation, uncertainty.*

Thèse de doctorat en Informatique, Laboratoire de Génie Informatique et d'Ingénierie de Production, école doctorale Information Structures Systèmes, Université de Montpellier, sous la direction de Jacky Montmain (Pr, IMT Mines Alès)et Patrice Bellot (Pr, Aix-Marseille Université, Laboratoire des Sciences de l'Information et des Systèmes, UMR 7296). Thèse soutenue le 23/11/2017.

Jury : M. Jacky Montmain (Pr, IMT Mines Alès, codirecteur), M. Patrice Bellot (Pr, Aix-Marseille Université, Laboratoire des Sciences de l'Information et des Systèmes, UMR 7296, codirecteur), Mme Catherine Berrut (Pr, Université de Grenoble, rapporteur), M. Pierre Zweigenbaum (DR, CNRS, LIMSI, Orsay, rapporteur), Mme Béatrice Daille (Pr, Université de Nantes, présidente), M. Mathieu Roche (MC, CIRAD, examinateur), M. Sébastien Harispe (MC, IMT Mines Alès, encadrant), Mme Sylvie Ranwez (Pr, IMT Mines Alès, encadrant).

Résumé : *Les concepts de découverte et d'extraction de connaissances ainsi que d'inférence sont abordés sous différents angles au sein de la littérature scientifique. En effet, de nombreux domaines s'y intéressent allant de la recherche d'information, à l'implication textuelle en passant par les modèles d'enrichissement automatique des bases de connaissances. Ces concepts suscitent de plus en plus d'intérêt à la fois dans le monde académique et industriel favorisant le développement de nouvelles méthodes.*

Cette thèse propose une approche automatisée pour l'inférence et l'évaluation de connaissances basée sur l'analyse de relations extraites automatiquement à partir de textes. L'originalité de cette approche repose sur la définition d'un cadre tenant compte (i) de l'incertitude linguistique et de sa détection dans le langage naturel, réalisée au travers d'une méthode d'apprentissage tenant compte d'une représentation vectorielle spécifique des phrases, (ii) d'une structuration des objets étudiés (p. ex. syntagmes nominaux) sous la forme d'un ordre partiel tenant compte à la fois des implications syntaxiques et d'une connaissance a priori formalisée dans un modèle de connaissances de type taxonomique (iii) d'une évaluation des relations extraites et inférées grâce à des modèles de sélection exploitant une organisation hiérarchique des relations considérées. Cette organisation hiérarchique permet de distinguer différents critères en mettant en oeuvre des règles de propagation de l'information permettant ainsi d'évaluer la croyance qu'on peut accorder à une relation en tenant compte de l'incertitude linguistique véhiculée.

Bien qu'a portée plus large, notre approche est ici illustrée et évaluée au travers de la définition d'un système de réponse à un questionnaire, généré de manière automatique, exploitant des textes issus du Web. Nous montrons notamment le gain informationnel apporté par la connaissance a priori, l'impact des modèles de sélection établis et le rôle joué par l'incertitude linguistique au sein d'une telle chaîne de traitement. Les travaux sur la détection de l'incertitude linguistique et la mise en place de la chaîne de traitement ont été validés par plusieurs publications et communications nationales et internationales. Les travaux développés sur la détection de l'incertitude et la mise en place de la chaîne de traitement sont disponibles au téléchargement à l'adresse suivante : https://github.com/PAJEAN/.

URL où le mémoire peut être téléchargé :
https://www.researchgate.net/publication/323458149_Gestion_de_l'incertitude_et_de_l'imprecision_dans_un_processus_d'extraction_de_connaissances_a_partir_des_textes

---

Rachel PANCKHURST : rachel.panckhurst@univ-montp3.fr

Titre : Entre linguistique et informatique. Des outils de traitement automatique du langage naturel écrit (TALNE) à l'analyse du discours numérique médié (DNM)

Mots-clés : Traitement automatique du langage naturel écrit (TALNE) en français, analyse de discours numériques médiés (courriels, forums, chats, SMS, WhatsApp, messageries instantanées), dispositifs novateurs en eLearning, évaluation de logiciels.

Title: *Between Linguistics and Computational Linguistics. From Written Natural Language Processing Tools to Mediated Digital Discourse Analysis*

Keywords: *Written natural language processing in French, mediated digital discourse analysis (email, forums, chats, SMS, WhatsApp, instant messaging), innovative eLearning methods, software evaluation.*

Habilitation à diriger des recherches en Informatique, Université Paris-Est, sous la direction de Panayota Kyriacopoulou (Pr, Université Paris-Est Marne-la-Vallée). Habilitation soutenue le 30/05/2017.

Jury : M. Georges Antoniadis (Pr, Université Grenoble-Alpes, examinateur), M. Cédrick Fairon (Pr, Université catholique de Louvain, Belgique, rapporteur), Mme Cvetana Krstev (Pr, Université de Belgrade, Serbie, examinatrice), Mme Panayota Kyriacopoulou (Pr, Université Paris-Est Marne-la-Vallée, directrice), M. Éric Laporte (Pr, Université Paris-Est Marne-la-Vallée, rapporteur), Mme Claudine Moïse (Pr, Université Grenoble-Alpes, examinatrice), M. Mathieu Roche (chercheur HDR, Cirad, UMR TETIS, Montpellier, examinateur), Mme Frédérique Segond (directrice du centre de R&D, Viseo, Grenoble, rapporteur).

Résumé : *Mon habilitation à diriger des recherches, intitulée « Entre linguistique et informatique. Des outils de traitement automatique du langage naturel écrit (TALNE) à l'analyse du discours numérique médié (DNM) », soutenue à la COMUE Université Paris-Est, se décline en trois volumes : volume I (synthèse), volume II (publications), volume III (curriculum vitæ)*[1]*. Ce résumé ne concerne que le volume I (synthèse).*

Depuis mon doctorat et ma nomination en tant qu'enseignante-chercheuse à l'université Paul-Valéry Montpellier 3 (en octobre 1992), mes activités d'enseignement, d'administration et de recherche s'inscrivent dans le domaine du traitement automatique du langage et des langues (TAL), et, plus précisément, du traitement automatique du langage naturel écrit (TALNE). Trois cheminements, ou volets de recherche, traversent et s'imbriquent tout au long de mes 25 années de carrière universitaire, jusqu'à présent :

1) prototypes et outils : interrogatives, verbes, gloses (1991-2003) ;

2) formation, (auto)évaluation, réseaux pédagogiques (technologies de l'information et de la communication éducatives (TICE), formation ouverte et à distance (FOAD, eLearning*)) (1996-2012) ;*

3) communication médiée par ordinateur (CMO), discours électronique médié (DEM), DNM : analyse de courriels, forums, chats, SMS (1996-2017).

Ceux-là sont explorés tout au long de ce manuscrit. La façon dont la recherche s'imprègne et s'enrichit de mes activités d'enseignement et d'administration est, je crois, cruciale. De ce fait, je présente l'ensemble de mes activités tripartites en tant qu'enseignante-chercheuse, afin que le lecteur puisse mieux entrevoir mon parcours global.

1. Les volumes II et III sont disponibles sur demande à l'adresse électronique indiquée ci-dessus.

À travers mon parcours, certes atypique — dans la mesure où je n'ai découvert l'espace francophone au quotidien qu'à partir de 18 ans —, j'espère pouvoir montrer comment j'ai contribué en recherche (mais également en pédagogie et en administration) au domaine de la linguistique informatique et, par conséquent, de quelle manière j'estime être en mesure d'animer des recherches doctorales. Dans une première section, je dessine brièvement mon parcours initial jusqu'au doctorat — afin de guider le lecteur à travers mes tout premiers pas de jeune chercheuse. La deuxième section est consacrée à l'évocation de mes activités d'enseignement et de formation et de mes responsabilités pédagogiques et administratives. Cela n'est peut-être pas habituel dans le cadre d'une habilitation, mais je m'octroie le droit de le faire, dans la mesure où je souhaite mettre en lumière leur importance pour moi et indiquer comment elles ont nourri ma réflexion en recherche. La troisième section constitue le « noyau dur » du manuscrit : la recherche. J'explique comment j'ai tissé les fils des volets et des thématiques, comment j'ai tâtonné, bifurqué, au gré des rencontres scientifiques. Puis, je montre aussi comment l'enseignement, l'administration et la recherche s'imbriquent, de manière plus approfondie. Dans la quatrième et dernière section, je mentionne les horizons et les perspectives à venir, avant de proposer une sélection globale de mes publications. Parmi les directions futures citées et les points clefs que j'estime fondamentaux à retenir dans un cadre de recherche publique, je les énumère ainsi en anglais dans ma conclusion — peut-être pourrait-on m'accuser d'être utopique ?

1) Deliver crucial research information to the general public *(fournir au grand public des informations de recherche cruciales).*

2) Request cross-disciplinary PhDs *(demander des doctorats interdisciplinaires).*

3) Demand that research results be factored into Ministerial reforms *(exiger que les résultats de la recherche soient pris en compte dans les réformes ministérielles).*

4) Help provide scientific expertise for devising real-life applications and software *(aider à fournir une expertise scientifique afin de concevoir des applications et des logiciels).*

5) Continue applied research—including PhD supervision—and help link up academic institutions with other organisations. *(poursuivre la recherche appliquée — y compris dans le cadre doctoral — et aider à établir des réseaux entre les établissements universitaires et d'autres organisations).*

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-01646172/

Marie-Sophie PAUSÉ : pauselinguist@gmail.com

Titre : Structure lexico-syntaxique des locutions du français et incidence sur leur combinatoire

Mots-clés : Locution, lexicologie, phraséologie du français, interface sémantique-syntaxe, Lexicologie Explicative et Combinatoire (LEC), Réseau Lexical du Français (RL-fr).

Title: *Impact of Lexico-Syntactic Structure of French Idioms on their Combinatory*

Keywords: *Idiom, lexicology, phraseology of the French language, semantic-syntax interface, Explanatory Combinatorial Lexicology (ECL), French Lexical Network (fr-LN).*

Thèse de doctorat en Sciences du langage, ATILF (CNRS et Université de Lorraine), Nancy, sous la direction de Alain Polguère (Pr, Université de Lorraine) et Sylvain Kahane (Pr, Université Paris Ouest – Paris 10). Thèse soutenue le 03/11/2017.

Jury : M. Alain Polguère (Pr, Université de Lorraine, codirecteur), M. Sylvain Kahane (Pr, Université Paris Ouest – Paris 10, codirecteur), M. Xavier Blanco (Pr, Universitat Autònoma de Barcelona, Espagne, rapporteur), Mme Christiane Fellbaum (Senior Research Scholar, Princeton University, États-Unis, rapporteur), Mme Éva Buchi (DR, CNRS, ATILF, Nancy,), Mme Agata Savary (MC HDR, Université François Rabelais de Tours, examinatrice).

Résumé : *En tant que syntagmes sémantiquement non compositionnels, les locutions sont des unités lexicales à part entière, qui doivent avoir leur propre entrée dans un modèle du lexique. Elles doivent donc recevoir une définition lexicographique, ainsi qu'une description de leurs caractéristiques grammaticales. De plus, en vertu de leur signifiant syntagmatique, les locutions témoignent — à des degrés divers — d'une flexibilité formelle (passivation, insertion de modificateurs, substitution de certains constituants, etc.).*

Cette thèse défend l'idée selon laquelle une description des locutions combinant à la fois l'identification des unités lexicales qui les composent et l'identification des relations de dépendance syntaxique qui unissent ces unités lexicales, permettra de prédire leurs différents emplois possibles dans la phrase. Une telle description n'est possible que dans un modèle du lexique décrivant précisément la combinatoire des lexies. Notre recherche, basée sur les principes de la Lexicologie Explicative et Combinatoire, exploite et enrichit les données du Réseau Lexical du Français (RL-fr), ressource en cours de développement à l'ATILF.

La thèse a deux principaux apports. Le premier est le développement d'un modèle de description lexico-syntaxique relativement fine des locutions du français. Le second est l'identification et l'étude de différentes variations structurales, syntaxiques et lexicales liées à la flexibilité formelle des locutions. Les variations des locutions sont mises en corrélation avec leurs structures lexico-syntaxiques, mais également avec

leurs définitions lexicographiques. Ceci nous conduit à introduire la notion de projection structurale, centrale dans le continuum de la flexibilité formelle des locutions.

La thèse est structurée en cinq chapitres, dont une introduction générale et une conclusion générale. L'introduction générale présente l'objet d'étude, les objectifs de la thèse, et la méthodologie employée. La notion de locution est présentée sous l'angle lexicographique, ce qui implique une mise en lumière des limites des modélisations lexicographiques des locutions françaises qui précèdent la modélisation préconisée dans le cadre de ce travail, à savoir celle du Réseau Lexical du français.

Le chapitre second a pour objectif, d'une part, de présenter les notions de base, et, d'autre part, de développer les caractéristiques des locutions en les positionnant par rapport aux autres classes de phrasèmes.

Le troisième chapitre présente les principes de description lexicographique des locutions que ce travail a permis d'établir dans le cadre du développement du Réseau lexical du français. Les locutions sont des unités lexicales à part entière, qui constituent des nœuds du réseau, au même titre que les lexèmes. Lexèmes et locutions appartiennent à la catégorie des lexies. La description lexicographique des locutions inclut des caractéristiques grammaticales, une définition, et des liens paradigmatiques et syntagmatiques avec d'autres nœuds du réseau. Parmi les caractéristiques grammaticales figure une structure lexico-syntaxique. Cette dernière permet d'identifier le patron syntaxique et les unités lexicales constituantes des locutions.

Le quatrième chapitre de la thèse présente le produit de la description lexico-syntaxique des locutions obtenu au terme de nos trois années de travaux — soit 498 patrons pour 2 821 locutions. Les patrons syntaxiques correspondent chacun à une structure syntaxique de surface, mais sont présentés sous format linéaire. Les positions actancielles contrôlées par certaines locutions sont prises en considération, et donnent lieu à des patrons spécifiques. La classification des patrons est opérée relativement aux types grammaticaux de locutions dénombrés (locutions verbales, locutions nominales, etc.). Le nombre de locutions associées à chaque patron est indiqué.

Reprenant la distinction opérée, au second chapitre, entre flexibilité formelle et défigement d'une locution, le chapitre cinq propose une classification puis une modélisation des variations formelles des locutions à partir d'exemples attestés. La dernière section du chapitre est consacrée à la description des variations formelles de 47 locutions verbales construites sur un patron syntaxique linéarisé du type `V Art NC`*. Les variations formelles suivantes sont étudiées : passivation, clivage, relativisation, variabilité du déterminant du constituant nominal, et attachement d'un dépendant syntaxique à un constituant autre que la tête de syntagme. Une modélisation à l'interface entre sémantique et syntaxe est proposée, sous l'angle de la projection structurale. Cette notion est introduite afin de rendre compte des correspondances entre des sémantèmes du réseau sémantique de la définition de la locution et des constituants de sa structure lexico-*

syntaxique. Les correspondances identifiées permettent d'activer certaines variations formelles.

URL où le mémoire peut être téléchargé :
http://hal.archives-ouvertes.fr/tel-01657880