Résumés de thèses et HDR

Rubrique préparée par Sylvain Pogodalla

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

Karën FORT : karen.fort@loria.fr

Titre : Myriadisation et éthique pour le traitement automatique des langues

Mots-clés : traitement automatique des langues, myriadisation, éthique, jeux ayant un but.

Title: *Crowdsourcing and Ethics for Natural Language Processing*

Keywords: *natural language processing, crowdsourcing, ethics, games with a purpose.*

Habilitation à diriger des recherches en informatique, LORIA, UMR 7503, sous la direction de Mme Claire Gardent (DR, CNRS). Habilitation soutenue le 23/11/2022.

Jury : Mme Claire Gardent (DR, CNRS, directrice), M. Jean-Yves Antoine (Pr, Université François Rabelais, Tours, rapporteur), Mme Iris Eshkol-Taravella (Pr, Université Paris Nanterre, rapporteuse), M. Dirk Hovy (associate professor, Bocconi University, Milan, Italie, rapporteur), M. Philippe Blache (DR, CNRS, président), Mme Armelle Brun (Pr, Université de Lorraine, examinatrice), M. Massimo Poesio (Pr, Université Queen Mary, Londres, Royaume-Uni, examinateur), Mme Marta Severo (Pr, Université Paris Nanterre, examinatrice).

Résumé : *Le traitement automatique des langues (TAL) a subi deux révolutions ces dix dernières années : le raccourcissement extrême de la distance entre les productions de la recherche et l'utilisateur final et l'avènement de l'apprentissage profond (*deep learning*). En conséquence, les besoins en données ont explosé en parallèle des questions éthiques. Cette habilitation à diriger des recherches présente les travaux que j'ai menés dans le domaine de la production d'annotations manuelles pour le TAL par myriadisation (*crowdsourcing*), en particulier par le jeu (*games with a purpose*),*

et dans celui de l'éthique pour le TAL. J'y redéfinis la myriadisation et les sciences participatives en général et je présente en détail les jeux ayant un but, leurs atouts et leurs limites. Je m'attarde plus particulièrement sur ZombiLingo, qui a servi à collecter des annotations en syntaxe de dépendances pour le français et RigorMortis, un jeu d'annotation d'unités polylexicales. Je me concentre dans une dernière partie sur l'éthique pour le TAL, un sous-domaine qui n'a véritablement été reconnu qu'à partir de 2016 et dont j'ai été précurseure. Je reviens sur son historique, son évolution récente et présente mes travaux, menés dans une optique plus déontologiste que conséquentialiste, permettant d'avoir une vision systémique du TAL et des problèmes éthiques qu'il pose.

URL où le mémoire peut être téléchargé :
https://theses.hal.science/tel-03873000

Jade MEKKI : jade.mekki@gmail.com

Titre : Caractérisation de registres de langue par extraction de motifs séquentiels émergents

Mots-clés : registres de langues, traitement automatique des langues, motifs séquentiels.

Title: *Characterisation of Language Registers Using Emerging Sequential Pattern Extraction*

Keywords: *language registers, natural language processing, sequential patterns.*

Thèse de doctorat en informatique, Expression, IRISA, UMR 6074, Université de Rennes 1, sous la direction de M. Damien Lolive (MC HDR, Université de Rennes 1), Mme Delphine Battistelli (Pr, Université de Paris Nanterre), M. Gwénolé Lecorvé (chercheur, Orange) et M. Nicolas Béchet (MC, Université de Bretagne-Sud). Thèse soutenue le 08/09/2022.

Jury : M. Damien Lolive (MC HDR, Université de Rennes 1, codirecteur), Mme Delphine Battistelli (Pr, Université de Paris Nanterre, codirectrice), M. Gwénolé Lecorvé (chercheur, Orange, codirecteur), M. Nicolas Béchet (MC, Université de Bretagne-Sud, codirecteur), Mme Farah Benamara (MC HDR, Université Paul Sabatier, rapporteuse), M. Thierry Charnois (Pr, Université Paris 13 Nord, rapporteur), M. Jean-Yves Antoine (Pr, Université François Rabelais, Tours, président), M. Olivier Baude (Pr, Université de Paris Nanterre, examinateur), M. Dominique Legallois (Pr, Université Sorbonne Nouvelle, examinateur).

Résumé : *Le locuteur d'une langue sent que pour un même message il existe plusieurs manières de le dire. Ce phénomène linguistique est celui des registres de langue. Ils renvoient à un trait saillant du langage que tout locuteur saisit intuitivement.*

Cette thèse s'intéresse à la caractérisation automatique des registres. Notre approche se fonde sur un large corpus de textes, considère divers niveaux d'analyse de la langue en même temps et caractérise les registres de manière comparative.

Sur le plan linguistique, notre contribution est d'étudier les apports des techniques de traitement automatique des langues pour extraire de nouvelles connaissances à propos des registres familier, courant et soutenu. Sur le plan informatique, nous avons proposé une méthode suffisamment générique et non supervisée pour caractériser tout type de variation linguistique, les registres s'apparentant alors à un cas d'usage.

Dans le manuscrit, nous dressons tout d'abord un état des lieux des multiples définitions présentes dans la littérature, par rapport auquel nous positionnons nos travaux. En effet, si les registres de langue semblent être un phénomène intuitivement reconnaissable et aisé à saisir, il n'existe aucun consensus sur leur définition dans la littérature scientifique.

Nous présentons ensuite la constitution linguistiquement motivée d'un large corpus de tweets en français étiquetés en registres. Les étiquettes découlent d'un procédé semi-supervisé fondé sur une graine annotée manuellement en registres et un classifieur qui généralise les annotations à l'ensemble des tweets. Le corpus étiqueté en résultant compte 228 505 tweets pour un total de 6 millions de mots.

À partir de ce corpus étiqueté, nous montrons que l'emploi de techniques d'extraction de motifs séquentiels émergents permet d'extraire des traits linguistiques caractéristiques des registres étudiés. Notre approche lève les trois principaux verrous de la fouille de motifs : la complexité algorithmique, l'abondance des motifs extraits et la difficulté d'évaluer ces derniers.

Le premier verrou est levé avec une méthodologie s'appuyant sur un ensemble minimal de traits linguistiques pour décrire chaque mot mais dont les croisements maximisent la description. La seconde limite est endiguée en réduisant le nombre de motifs extraits en les partitionnant automatiquement. Enfin, le dernier verrou est désamorcé en proposant deux protocoles d'évaluations indépendantes (automatique et perceptuelle).

Le manuscrit s'achève avec l'application de notre approche à une autre problématique : caractériser différents genres de textes adressés aux enfants. Les résultats obtenus indiquent que notre approche est robuste et peut être généralisée à d'autres cas d'usage.

URL où le mémoire peut être téléchargé :
https://hal.science/tel-03991094

Filip MILETIC : filip.miletic@ims.uni-stuttgart.de

Titre : Étude des glissements de sens induits par le contact de langues en anglais québécois : apports conjoints de la modélisation vectorielle sur corpus et de l'approche sociolinguistique variationniste

Mots-clés : glissements de sens, anglais québécois, modèles sémantiques vectoriels, corpus de tweets, sociolinguistique variationniste, contact de langues.

Title: *An Investigation into Contact-Induced Semantic Shifts in Quebec*

Keywords: *semantic shifts, Quebec English, vector space models, Twitter corpora, variationist sociolinguistics, language contact.*

Thèse de doctorat en sciences du langage, CLEE, Université Toulouse - Jean Jaurès, sous la direction de Mme Anne Przewozny-Desriaux (Pr, Université Toulouse - Jean Jaurès) et M. Ludovic Tanguy (MC, Université Toulouse - Jean Jaurès). Thèse soutenue le 20/06/2022.

Jury : Mme Anne Przewozny-Desriaux (Pr, Université Toulouse - Jean Jaurès, codirectrice), M. Ludovic Tanguy (MC, Université Toulouse - Jean Jaurès, codirecteur), M. Stefan Dollinger (Pr, University of British Columbia, Vancouver, Canada, rapporteur), Mme Sabine Schulte im Walde (Pr, Universität Stuttgart, Allemagne, rapporteuse, présidente), M. Kris Heylen (researcher, KU Leuven, Louvain, Belgique, examinateur), Mme Amélie Josselin-Leray (MC, Université Toulouse - Jean Jaurès, examinatrice).

Résumé : *Cette thèse étudie les glissements de sens induits par le contact de langues en anglais québécois, à savoir des mots anglais préexistants utilisés avec un sens différent en raison d'une influence potentielle du français. Ce phénomène sociolinguistique est décrit dans plusieurs études antérieures, mais il reste de nombreuses inconnues quant à sa diffusion, aux contraintes sur ses usages et à la valeur sociale qu'il véhicule. Nous proposons une approche novatrice à l'intersection du traitement automatique des langues et de la sociolinguistique variationniste, afin de fournir une description exhaustive de ce phénomène ainsi que d'évaluer les contributions des approches sur corpus mises en œuvre ici.*

Afin d'effectuer des analyses computationnelles de variation sémantique, nous avons constitué un corpus composé de 78.8 millions de tweets, publiés par 196 000 locuteurs de Montréal, Toronto et Vancouver. Le corpus a été utilisé pour mettre en œuvre différents types de modèles vectoriels, à savoir des représentations computationnelles du sens des mots. Les modèles statiques ont permis d'identifier de nouveaux glissements de sens (en identifiant des différences entre les locuteurs de Montréal par rapport aux deux autres villes), alors que les modèles contextuels ont permis de caractériser plus finement leurs utilisations. Malgré des résultats prometteurs, les analyses qualitatives indiquent que ces méthodes sont limitées par le bruit lié à leurs caractéristiques intrinsèques et à la structure du corpus. Ceci est corroboré par une évaluation quantitative systématique effectuée sur un jeu de données composé de 80 items. Celle-ci a montré

que des résultats comparables à l'état de l'art sur une tâche classique de détection de changement sémantique ne se traduisent pas directement par la capacité pratique à repérer de nouveaux glissements de sens.

Ces approches à grande échelle ont été complétées par des données plus fines recueillies au moyen d'entretiens sociolinguistiques avec 15 locuteurs vivant à Montréal. Nous avons utilisé un protocole sociophonologique classique, garantissant des résultats comparables et fiables, ainsi qu'un nouveau test de perception portant sur l'acceptabilité de 40 glissements de sens attestés dans le corpus de tweets. Les corrélations entre ces variables linguistiques et différents facteurs sociodémographiques, ainsi que les remarques qualitatives sur leur utilisation, indiquent quatre patterns de variation synchronique. Ceux-ci pourraient à leur tour refléter des processus diachroniques. Par ailleurs, la variabilité inter-locuteurs suggère un rôle important des locuteurs bilingues et plus jeunes dans l'utilisation des glissements de sens. Enfin, les scores d'acceptabilité sont faiblement corrélés avec les mesures computationnelles, ce qui suggère que ceux-ci reflètent d'autres dimensions de variation sémantique.

Dans l'ensemble, cette thèse a fourni la première description systématique, menée sur corpus et au moyen d'entretiens, des glissements de sens en anglais québécois induits par le contact avec le français. Elle a également mis en évidence la complémentarité des approches développées dans des disciplines différentes : notre objet d'étude sociolinguistique a orienté la mise en place des expériences computationnelles ; celles-ci ont fourni les stimuli utilisés dans les entretiens sociolinguistiques ; ces derniers ont apporté une évaluation supplémentaire des méthodes computationnelles. Ces considérations ouvrent la voie à une utilisation plus avisée des méthodes computationnelles basées sur corpus dans des études de phénomènes sociolinguistiques.

URL où le mémoire peut être téléchargé :
https://dante.univ-tlse2.fr/s/fr/item/32205
