Résumés de thèses et HDR

Rubrique préparée par Sylvain Pogodalla

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

Gabriele CHIGNOLI : gabrielechignoli@icloud.com

Titre : Les composantes de la parole dans la caractérisation phonétique du locuteur : étude sur la complémentarité et la redondance véhiculées des informations

Mots-clés : locuteur, caractéristiques, composantes, CNN, spontané, clustering, informativité, comparaison.

Title: *Speech Components in Phonetic Characterisation of Speakers: A Study on Complementarity and Redundancy of Conveyed Information*

Keywords: *speaker, characteristics, components, CNN, spontaneous, clustering, informedness, comparison.*

Thèse de doctorat en sciences du langage, Laboratoire de Phonétique et Phonologie, UMR 7018, Université Sorbonne Nouvelle, sous la direction de M. Cédric Gendrot (Pr, Université Sorbonne Nouvelle). Thèse soutenue le 15/09/2022.

Jury : M. Cédric Gendrot (Pr, Université Sorbonne Nouvelle, directeur), M. Damien Lolive (Pr, Université de Rennes, IRISA, rapporteur), Mme Ioana Vasilescu (CR HDR, CNRS, LISN, rapporteuse), Mme Cécile Fougeron (DR, CNRS, Laboratoire de Phonétique et Phonologie, UMR 7018, examinatrice), M. Jean-François Bonastre (Pr, Université d'Avignon, examinateur), Mme Christine Meunier (CR HDR, CNRS, présidente).

Résumé : *The decomposition of the speech signal into phonetically meaningful units allows the analysis of between- and within- speaker variations. These are components associated with characteristics whose nature relates to the physical, psychological and social aspects of a speaker. In this thesis, we compare perceptual characterisation results with a phonetic analysis and advanced modelling techniques through*

Convolutional Neural Networks (CNN). Two French corpora of read and spontaneous speech are the object of our studies from which some results have emerged that should be considered important for the speaker characterisation domain.

The characteristics allowing the description of variation of speech components occurring in different stimuli by a single speaker appear consistent with the phonetic measurements that play an important role in the separation of stimuli by different speakers. This allows creating multiple groups of speakers inside the studied population that are characterised by similar distributions of speech components. In this sense, we observe that source and filter characteristics are more important in the description of female speakers' variation, while voice quality characteristics such as breathiness and hoarseness have a greater impact on male speakers. This suggests that the characterisation of female speakers relies more on linguistic and articulation aspects, as well as the role of interlocutors in the conversation, i.e., the example of formant dispersion, whereas paralinguistic aspects, such as the level of confidentiality between speakers that changes in breathiness may convey, are retained for the characterisation of male speakers.

The perceptual responses confirm these tendencies, with the human-based clustering showing consistency with CNN- and phonetic-based results. In particular, the clustering analysis further highlights consistency of CNN results with the statistical analysis of speech components, supporting further application of these methods for phonetic studies.

The last highlight of this thesis concerns the role of Mel Frequency Cepstral Coefficients (MFCC) in comparison to classical phonetic measurements. These show a great adaptation to speakers' characteristics, relating to different aspects for female and male speakers, and for the multiple groups of speakers present in our population. Rather than being representative of a specific trait, MFCC are mainly linked to intensity and fundamental frequency for female speakers characterisation, while to the distributions of energy and low-level spectral shape for male speakers.

URL où le mémoire peut être téléchargé :
https://hal.science/tel-03911819

Ghazi FELHI : ghazi.felhi@gmail.com

Titre : Représentations de phrases interprétables avec autoencodeurs variationnels et attention

Mots-clés : autoencodeurs variationnels, transformeurs, interpretabilité, désenchevêtrement, apprentissage semi-supervisé, apprentissage non supervisé, modèle de langue, syntaxe.

Title: *Interpretable Sentence Representation with Variational Autoencoders and Attention*

Keywords: *variational autoencoders, transformers, interpretability, disentanglement, semi-supervised learning, unsupervised learning, language modeling, syntax.*

Thèse de doctorat en informatique, Laboratoire d'Informatique de Paris-Nord, Université Sorbonne Paris-Nord, sous la direction de Mme Adeline Nazarenko (Pr, Université Sorbonne Paris-Nord), M. Joseph Le Roux (MC, Université Sorbonne Paris-Nord) et M. Djamé Seddah (MC, Université Paris Sorbonne, Inria). Thèse soutenue le 26/02/2023.

Jury : Mme Adeline Nazarenko (Pr, Université Sorbonne Paris-Nord, codirectrice), M. Joseph Le Roux (MC, Université Sorbonne Paris-Nord, codirecteur), M. Djamé Seddah (MC, Université Paris Sorbonne, Inria, codirecteur), M. Benjamin Piwowarski (CR, CNRS, Institut des Systèmes Intelligents et de Robotique, rapporteur), M. François Yvon (DR, CNRS, LISN, rapporteur), M. Laurent Besacier (Pr, Université Grenoble Alpes, président).

Résumé : *Dans cette thèse, nous développons des méthodes pour améliorer l'interprétabilité de techniques récentes d'apprentissage de représentation en traitement automatique de langues (TAL) en prenant en compte la difficulté d'obtention de données annotées. Nous utilisons des autoencodeurs variationnels (VAE) afin d'apprendre avec peu de données des représentations interprétables. Pour notre première contribution, nous identifions et supprimons des composants inutiles du fonctionnement des VAE semi-supervisés, améliorant ainsi leur vitesse de calcul et facilitant leur conception. Notre deuxième et principale contribution consiste à utiliser des VAE et des transformeurs pour construire deux modèles qui permettent de séparer l'information dans les représentations latentes en concepts interprétables sans données annotées. Le premier modèle, ADVAE, est capable de représenter et de contrôler séparément des informations sur les rôles syntaxiques dans les phrases. Le second modèle, QKVAE, utilise des variables latentes séparées pour former des clés et des valeurs pour son décodeur transformeur et est capable de séparer les informations syntaxiques et sémantiques dans ses représentations neuronales. Dans des expériences de transfert, QKVAE a une performance compétitive par rapport aux modèles supervisés et une performance équivalente à un modèle supervisé utilisant 50 000 échantillons annotés. De plus, QKVAE montre une capacité améliorée de désenchevêtrement des rôles syntaxiques par rapport à ADVAE. De manière générale, notre travail montre qu'il est*

possible d'améliorer l'interprétabilité des architectures de pointe utilisées pour les modèles de langage avec des données non annotées.

URL où le mémoire peut être téléchargé :
https://arxiv.org/abs/2305.02810

Laura NORESKAL : laura.noreskal@outlook.fr

Titre : Erreurs dans les phrases coordonnées au sein des rédactions universitaires : typologie et détection

Mots-clés : erreurs, rédactions universitaires, détection automatique de l'erreur, phrase coordonnée, linguistique de corpus, TAL, classification supervisée.

Title: *Errors in Coordinated Sentences in Academic Writing: Typology and Detection*

Keywords: *errors, students writings, error detection, coordinated sentence, corpus linguistics, NLP, supervised classification.*

Thèse de doctorat en sciences du langage, MoDyCo, UMR 7114, UFR Phillia, Université Paris Nanterre, sous la direction de Mme Iris Eshkol-Taravella (Pr, Université Paris Nanterre) et de Mme Marianne Desmets (MC, Université Paris Nanterre). Thèse soutenue le 14/12/2022.

Jury : Mme Iris Eshkol-Taravella (Pr, Université Paris Nanterre, codirectrice), Mme Marianne Desmets (MC, Université Paris Nanterre, codirectrice), Mme Anne Abeillé (Pr, Université Paris Cité, présidente), Mme Frédérique Sitri (Pr, Université Paris-Est Créteil, rapporteuse), M. Olivier Kraif (Pr, Université Grenoble Alpes, rapporteur), Mme Sarah de Vogüé (MC, Université Paris Nanterre, examinatrice), Mme Silvia Adler (Pr, Université Bar-Ilan, Israël, examinatrice), M. Éric Villemonte de la Clergerie (CR, Inria, examinateur).

Résumé : *Face aux difficultés rédactionnelles rencontrées par les étudiants à leur entrée dans l'enseignement supérieur, une quinzaine d'universités françaises ont décidé de se réunir pour proposer des solutions de remédiation dans le cadre d'un projet nommé écri+ (ANR 17-NCUN-0015). Le projet écri+ a pour but de permettre aux étudiants francophones d'améliorer leurs compétences langagières en leur proposant des outils d'évaluation, de formation et de certification pour l'expression et la compréhension écrite du français. Parmi les difficultés observées, on retrouve les constructions syntaxiques complexes et les séquences phrastiques longues, avec des coordinations ou des juxtapositions. Ainsi, afin que les étudiants puissent s'autoformer sur la reconnaissance de ce type d'erreurs dans leurs textes, écri+ propose de mettre à leur disposition un outil de détection automatique d'erreurs dans les phrases coordonnées. En mêlant TAL, didactique et linguistique de corpus, cette recherche porte sur l'étude et la détection automatique des erreurs dans les constructions coordonnées issues des rédactions des étudiants. Après avoir constitué le corpus de rédactions composé de*

mémoires, rapports de stage, exercices et devoirs maison, nous avons procédé à l'analyse manuelle des données afin d'élaborer une typologie des erreurs réalisées dans les phrases coordonnées. La recherche réalisée a montré que les erreurs sont les plus présentes dans les productions non préparées telles que les exercices, et sont conditionnées par la taille des phrases, mais également par le nombre de coordonnants présents dans la phrase. Ensuite, le corpus a été annoté selon une typologie proposée et a été exploité pour le développement de l'outil de la détection automatique de ces erreurs. Dans un premier temps, l'outil développé prédit la classe, correcte ou erronée, pour une phrase coordonnée donnée. Dans un second temps, l'outil catégorise l'erreur reconnue, c'est-à-dire qu'il classe l'erreur parmi les 11 types proposés.

URL où le mémoire peut être téléchargé :
https://www.theses.fr/s241290

Mathilde REGNAULT : regnaultm@icloud.com

Titre : Annotation et analyse syntaxique de corpus hétérogènes : le cas du français médiéval

Mots-clés : annotation syntaxique, métagrammaire, français médiéval, ancien français, corpus hétérogène, grammaire d'arbres adjoints, parsing.

Title: *Syntactic Analysis and Parsing of Heterogeneous Corpora: The Case of Medieval French*

Keywords: *syntactic annotation, metagrammar, Medieval French, Old French, heterogeneous corpus, tree-adjoining grammar, parsing.*

Thèse de doctorat en sciences du langage, Lattice, UMR 8094, UFR Littérature, Linguistique, Didactique, Université Sorbonne Nouvelle, sous la direction de Mme Sophie Prévost (DR, CNRS, Lattice) et de M. Éric Villemonte de la Clergerie (CR, Inria). Thèse soutenue le 16/06/2022.

Jury : Mme Sophie Prévost (DR, CNRS, Lattice, codirectrice), M. Éric Villemonte de la Clergerie (CR, Inria, codirecteur), M. Sylvain Kahane (Pr, Université Paris Nanterre, rapporteur), Mme Laura Kallmeyer (Pr, Heinrich Heine Universität Düsseldorf, Allemagne, rapporteuse), Mme Béatrice Daille (Pr, Université de Nantes, examinatrice), Mme Annie Forêt (MC, Université de Rennes 1, examinatrice), M. Achim Stein (Pr, Universität Stuttgart, Allemagne, examinateur).

Résumé : *Le français médiéval couvre les états de langue d'ancien français (IX^{e} – XIII^{e} siècle) et de moyen français (XIV^{e} – XV^{e} siècle). Nous disposons de données annotées pour ces états de langue, dont le SRCMF, un corpus arboré d'ancien français. Il est cependant difficile d'obtenir plus de données annotées syntaxiquement, car les spécialistes sont peu nombreux et il n'existe pas encore d'outil dédié pour l'ensemble de la période. Développer ce genre d'outil permet d'obtenir des annotations plus facilement et d'en contrôler la qualité. Cependant, ce n'est pas une tâche simple*

parce que les différents états de langue sont soumis à la variation, due à plusieurs facteurs, notamment l'absence de norme graphique, la variation dialectale, la souplesse de l'ordre des mots, l'évolution de la morphologie et de la syntaxe (sur sept siècles), qui fait passer le français d'une langue SOV à une langue SVO. La nature des écrits se diversifie aussi à mesure que la littérature évolue et que le latin est délaissé au bénéfice du français comme langue administrative et juridique. Les données à analyser sont donc hétérogènes, ce qui rend difficile le traitement automatique, comme l'ont précédemment montré des expériences d'annotation morphosyntaxique sur le SRCMF.

Pour obtenir un parseur du français médiéval, nous proposons d'adapter la métagrammaire du français contemporain FRMG. Bien que les différents états de langue présentent des différences manifestes, les points communs sont suffisants pour rendre possible la modification d'un système existant pour obtenir un outil dédié. Les changements concernent essentiellement l'ordre des mots (constituants majeurs, modifieurs du nom, position des pronoms conjoints). Pour utiliser cet outil sur corpus, il est nécessaire d'enrichir le lexique d'ancien français OFrLex, d'une part pour obtenir une couverture lexicale satisfaisante sur les textes, et, d'autre part, pour y intégrer des informations syntaxiques et sémantiques nécessaires à l'analyse syntaxique.

Le développement d'un parseur symbolique vient, d'une part, de la volonté de justifier linguistiquement des analyses, ce qui permet de confronter notre compréhension de la syntaxe du français médiéval aux sorties du parseur, et, ainsi, de l'affiner. D'autre part, nous souhaitons nous servir de la fouille d'erreurs pour améliorer les divers composants de la chaîne (segmenteur, lexique, grammaire). Ce système est encore en développement, mais il nous permet déjà d'annoter des données de toute la période du français médiéval et de comparer ces analyses à celles de parseurs neuronaux, ce qui participe à orienter le travail de relecture vers des exemples difficiles à traiter.

URL où le mémoire peut être téléchargé :
https://theses.fr/s195294