ORALIDIA : Oralité et diachronie : une voie d'accès au changement linguistique

Objectifs :

L'oral est souvent considéré comme le lieu privilégié du changement linguistique (Blanche-Benveniste 2002), ce qui peut expliquer l'intérêt pour l'étude des marques d'oralité, de la représentation de l'oral ou du lien langue parlée / langue écrite (Schnedecker & Vaguer (dir.) 2020, Glikman et al. (dir.) 2019, Ayres-Bennet et al. (dir.) 2018, Blasco & Bodelot (dir.) 2017, Rodriguez Somolinos (dir.) 2016, Lagorgette & Larrivée (dir.) 2013). Cependant, aujourd'hui encore, malgré le développement des corpus oraux (ORFEO, ESLO), l'accès à l'oral spontané reste difficile. Le projet ORALIDIA vient combler ce manque, à travers la constitution d'un corpus inédit de français parlé spontané : les « sms vocaux » ou « vocaux ». Ces données sont spontanément produites en dehors de toute enquête ou entretien linguistique, et constituent une voie d'accès à la parole spontanée non surveillée, nécessaire pour la description de la langue naturelle. Ces données sont ainsi le lieu privilégié d'étude de la diffusion des formes émergentes ou de leur disparition, et du français parlé dans différents contextes, en particulier informels. Ce projet sera également l'occasion de mettre en avant le patrimoine et les particularités linguistiques de la région Grand Est, en privilégiant le recueil auprès des populations de Nancy et Strasbourg. En effet, les corpus de référence existants sont constitué d'entretiens de locuteurs de la région parisienne (cf. CFPP2000) ou d'Orléans (cf. ESLO). Le partenariat avec l'Université de Liège permettra également d'élargir le recueil au français tel qu'il est parlé à Liège. La comparaison avec les corpus écrits existants (dont SMS, Panckhurst et al. 2013 ; Stark 2016-2018) et oraux permettra de mener des études sur les dia-variations contemporaines, en micro-diachronie, ou selon le type de communication (objectifs, planification, destinataire, médium employé, proximité ou distance communicative (Koch & Oesterreicher 2001) …), pour des études linguistiques comme phonétiques.

Ce projet profite de l'expertise développée durant la délégation de recherche de Julie Glikman (porteur) au laboratoire ATILF en 2021-2022, durant laquelle une étude exploratoire auprès de 7 locuteurs a permis de tester la faisabilité et l'intérêt scientifique du projet. Les premiers résultats montrent des caractéristiques bien connues pour la production de l'oral (faux départs, hésitations, pauses, etc., Blanche-Benveniste et al. 1990), et permettent d'analyser les marqueurs de discours (« du coup », « genre », « tu vois », « en mode ») et les éléments de structuration du discours (« voilà »). On trouve aussi bien des communications de type interactionnelle (1) que des récits, pouvant comporter du discours rapporté au style direct, difficilement observable dans les corpus existants (2) :

(1) […] et qu'est-ce qui t'a déclenché cette pensée _ explique-moi _ t'es pas moche _ je te le dis (06_03)
(2) […] et il dit bonjour alors on est là genre bonjour _ et euh _ il sort à #name euh _ ouais euh vous pensez à bien garder les distances de sécurité hein euh alors elle dit ben oui vous voyez bien qu'ils sont espacés _ oui oui non mais je vois je vois hein […] (05_08)

Déroulement et mise en œuvre du projet :

Le projet ORALIDIA se déroule en trois phases majeures : i. adaptation du protocole exploratoire et recueil des données ; ii. transcription, enrichissement par l'annotation, et analyse des données, en s'appuyant sur les conventions de transcription, le guide d'annotation et la grille d'analyse établis sur le corpus exploratoire (délégation J. Glikman 2021-22), et adaptés si besoin pour les nouvelles données ; iii. publication des résultats et réponses à appel à projets internationaux pour élargir le recueil au français d'Europe (France, Belgique, Suisse). Deux réunions plénières par an sont prévues.

Le recueil des sms vocaux à grande échelle demande la mise en place d'un protocole de protection des données personnelles, d'une procédure technique de récupération des messages, et, après le recueil, un important travail de tri, nettoyage, anonymisation, transcription et annotations des données avant son analyse et sa diffusion. Un appel à la foule (*crowdsourcing*) sera fait, ce qui demandera un travail de communication et diffusion de l'appel à participants.

Le protocole de protection des données personnelles sera élaborée avec la collaboration de la Déléguée à la protection des données de l'Université de Strasbourg. La procédure de récupération des données

nécessite l'achat d'un téléphone avec les fonctionnalités modernes de messagerie connectée (financement par le présent appel) et d'une ligne de téléphone (financement J. Glikman) et pourra être assisté par un ingénieur pour la mise en place technique. J. Glikman procédera au tri, nettoyage et anonymisation des données dans le respect de la procédure RGPD mise en place. Pour la transcription, un partenariat de recherche avec l'outil de transcription automatique Vocapia (https://www.vocapia.com/) est envisagé pour effectuer une reconnaissance vocale, et obtenir ainsi une première transcription automatique. Cette transcription devra faire l'objet d'une vérification et correction manuelle. Dans ce but, le recrutement de stagiaires en Master 2 de linguistique est nécessaire (financement par le présent appel). Le volet annotation sera réparti entre l'alignement au phonème (responsable C. Fauth), l'annotation en POS (responsable C. Benzitoun) et l'annotation syntaxique (responsable N. Mazziotta), grâce au recrutement de stagiaires et de vacataires (niveau doctorat) (financement par le présent appel). Le corpus sera ensuite mis dans un format accessible à la communauté, par ex. via le format du logiciel de textométrie TXM ou la plateforme MatchGrew. Le corpus final sera diffusé sur la plateforme Ortolang (EquiPex, https://www.ortolang.fr/). La valorisation du corpus et les premières analyses seront présentées dans des colloques internationaux et feront l'objet d'articles scientifiques (financement présent appel à projet et laboratoires des membres).

Résultats attendus et livrables :

Les livrables envisagés dans ce projet sont (a) la constitution d'un corpus de français parlé spontané, distribué librement sous licence CC et suivant les principes FAIR, (b) des actions de diffusion et valorisation de la recherche, à travers des communications dans des colloques internationaux (LREC, UD, JEP) et la rédaction d'articles scientifiques (revues *Corpus*, *Travaux de Linguistique*, *Langue Française*), (c) le montage de projets internationaux financés pour élargir le recueil au français d'Europe (ANR, projet Tournesol pour la coopération franco-belge, programme Interreg franco-suisse).

Prise de risque :

Ce projet prévoit le recueil et la manipulation de données inédites qui n'ont encore jamais fait l'objet d'une étude systématique ni d'un recueil en vue de constituer un corpus. Cela demande des précautions tant en terme d'éthique de la recherche (protection des données personnelles) qu'en terme de faisabilité (besoins techniques). L'appel à la foule ajoute également une inconnue à la quantité de données recueillies lors du projet : le positionnement géographique du porteur et des partenaires permettra d'inciter localement les participations (auprès du public étudiant notamment), mais par le « bouche à oreille », la participation peut s'amplifier et fournir un nombre important de sms vocaux recueillis (par ex. le projet sur les SMS écrits mené à Montpellier a permis de recueillir plus de 88 000 sms), nécessitant ainsi la mise en place d'une stratégie stable et robuste.

Caractère amont du projet :

Première étude d'une source de données inédites.

Effet de levier pour le site :

Ce projet placera l'Université de Strasbourg au centre d'un projet de grande envergure pouvant être étendu à la France entière et aux pays francophones frontaliers. Ainsi, l'expertise pourra être transposée à d'autres aires linguistiques. Le corpus, diffusé pour la communauté internationale, servira à tous les chercheurs s'intéressant à la langue française, à la variation géographique, aux spécificités de la langue orale et de la parole spontanée.

Établissement de preuves et premières bases pour répondre à des projets nationaux et internationaux :

Les premiers résultats obtenus dans le cadre de cet IdEx permettront de prouver la faisabilité d'un tel projet, son intérêt scientifique pour la recherche en sciences du langage, et la robustesse du protocole de recherche et des outils mis en œuvre, ainsi que l'expertise des membres du projet.

Strasbourg, le 14 janvier 2022

Julie GLIKMAN

Annexe : Bibliographie

Ayres Bennet, W., A. Carlier, J. Glikman, T. M. Rainsford, G. Siouffi et C. Skupien Dekens (dir.) (2018) *Nouvelles voies d'accès au changement linguistique*, Classiques Garnier. https://classiques-garnier.com/nouvelles-voies-d-acces-au-changement-linguistique.html

Blanche-Benveniste *et al.* (1990) *Le français parlé*, CNRS éditions.

Blanche-Benveniste, C. (2002). Quel est le rôle du français parlé dans les évolutions syntaxiques ? *L'information grammaticale*, *94*(1), 11-17.

Blasco, M. & C. Bodelot (dir.) (2017) « Langue parlée / langue écrite, du latin au français : un clivage dans l'histoire de la langue ? », *Langages* 208.

CFPP2000 : http://cfpp2000.univ-paris3.fr/

ESLO : http://eslo.huma-num.fr/index.php/pagecorpus/pageaccescorpus

Estade, L. & J. le Maire (2018) « Les messages vocaux sur WhatsApp: une forme hybride de communication. » *Cahiers du Centre de Linguistique et des Sciences du Langage*, 2018, no 55, p. 125-134.

Glikman, J., G. Parussa, R. Waltereit (dir.) (2019) *Les marqueurs du discours en diachronie du français : nouvelles perspectives. Studia Linguistica Romanica*, n° 2, 2019, https://studialinguisticaromanica.org/index.php/slr/issue/view/2.

Koch P. & W. Oesterreicher W. (2001) « Gesprochene Sprache und geschriebene Sprache / Langage parlé et langage écrit », in Holtus G., Metzeltin M., Schmitt Ch. (éds), *Lexikon der Romanistischen Linguistik*, Bd. I/2, Tübingen, Niemeyer, p. 584-627.

Lagorgette D. & P. Larrivée (dir.) (2013) *Représentations du sens linguistique 5*, Presses de l'Université de Savoie.

MatchGrew : http://match.grew.fr/

ORFEO : https://www.projet-orfeo.fr/

Panckhurst R., Détrie C., Lopez C., Moïse C., Roche M. et Verine B. (2013). « Sud4science, de l'acquisition d'un grand corpus de SMS en français à l'analyse de l'écriture SMS ». *Épistémè - revue internationale de sciences sociales appliquées*, 9 : Des usages numériques aux pratiques scripturales électroniques, 107-138.

Rodriguez Somolinos A. (dir.) (2016) *Énonciation et marques d'oralité dans l'évolution du français*, *Linx* 73, Presses universitaires de Paris Nanterre.

Schnedecker C. & C. Vaguer (dir.) (2020) *L'oral représenté en diachronie et en synchronie : une voie d'accès à l'oral spontané ?*, *Langages* 2020/1 (N° 217).

Stark, Elisabeth (2016-2018). Projet FSN « What's up, Switzerland? » (Sinergia: CRSII1_160714). Université de Zurich. www.whatsup-switzerland.ch.

TXM : https://txm.gitpages.huma-num.fr/textometrie/index.html