LangaJE 2024 –
Données et corpus en linguistique et en didactique à la lumière de la science ouverte : problématiques et enjeux méthodologiques

L'équipe interuniversitaire LangaJE, affiliée au Centre de recherche en linguistique appliquée (CeRLA – Université Lumière Lyon 2) et au Centre d'Études Linguistiques – Corpus, Discours et Sociétés (CEL – Université Jean Moulin Lyon 3), organise, le **vendredi 15 novembre 2024 à Lyon** (France), sa troisième JE pour doctorant.es et jeunes chercheur.euses sur le thème de la **collecte de données et la constitution de corpus au regard des problématiques soulevées par la science ouverte**.

La première édition de cette journée d'étude (18 novembre 2022) avait pour thématique la valorisation des données issues de la recherche hors de l'université, tandis que la deuxième édition (13 octobre 2023) avait pour fil rouge la place de l'interdisciplinarité dans l'approche linguistique. Au cours de nos dernières discussions, nous avons constaté que l'ouverture aux autres disciplines permet d'exploiter des jeux de données avec un regard complémentaire et/ou parfois novateur. Centrée sur les données (écrites ou orales), cette journée abordera ainsi des questions relatives à leur collecte, leur nature et/ou leur compilation.

Au-delà des questions d'accessibilité, de diffusion et de valorisation des jeux de données et des corpus, les enjeux, la représentativité, le contexte de leur collecte ou compilation ainsi que la qualité des métadonnées sont autant d'éléments qui peuvent donner ou retirer du crédit à la production du chercheur (Biber 1993, Léon 2008). Ces éléments doivent ainsi faire l'objet d'une réflexion en amont, en tout début de projet, tant pour le projet de recherche en lui-même que pour l'ouverture, la diffusion, voire la réutilisation de ces données.

Par ailleurs, l'ouverture de la science ne concerne pas seulement l'accessibilité des données, des études et leur valorisation. En effet, elle peut également signifier l'inclusion du grand public à différents égards. Par exemple, la participation citoyenne peut se trouver directement et volontairement dans la création de bases de données et de recherches, comme pour le *Corpus Français Parlé de nos Régions* (CFPR) ou le projet Lingscape au Luxembourg (Purschke 2017). En outre, les recherches ne concernent pas uniquement des langues largement parlées telles que le français ou l'anglais : elles permettent de travailler sur des variantes moins répandues, comme le suisse allemand avec la plateforme *gschmöis* (Hasse, Bachmann et Glaser 2019), ou sur des langues régionales pour lesquelles les données manquent, comme le breton (Jouitteau 2012) ou l'alsacien (Millour et Fort 2019).

Enfin, la perspective de la science ouverte amène également à considérer des corpus créés dans un but plus large que celui de la recherche : ce double objectif concerne fréquemment la recherche en didactique des langues et les corpus d'apprenant.es, qui visent à aider les enseignant.es dans l'enseignement-apprentissage des langues (projet E-CALM : corpus EcriScol, corpus RésolCo ; projet MUST : *Multilingual Student Translation Corpus*, décrit dans Granger et Lefer 2020), sans que cet objectif soit exclusif à ce domaine. Lors de cette journée d'étude, nous entendons donc questionner la manière de collecter des données et d'élaborer des corpus pensés en amont dans le cadre de la science ouverte, qu'ils soient destinés à être en libre accès, qu'ils aient un but applicatif, ou qu'ils soient créés grâce à la science citoyenne. Dans cette optique, et au regard des enjeux de la science ouverte, les communications pourront aborder :

- les outils, aspects méthodologiques et/ou théoriques concernant la collecte de données dans le cadre de la science participative ou dans le cadre de la collecte de données publiques à grande échelle ;

- les enjeux méthodologiques inhérents à la question des corpus (critères de constitution ; accessibilité ; exploration ; restrictions (Abouda & Baude 2006, Blanche-Benveniste 2007, Meunier 2008) ou théoriques (Condamines 2005) ;

- Toute réflexion concernant la nature des données collectées ou mobilisées dans la recherche en linguistique et en didactique des langues.

Ces trois axes ne sont pas exclusifs et des communications les associant sont les bienvenues. Par ailleurs, les éléments listés le sont à titre illustratif et nous encourageons toute proposition en lien avec la thématique de cette journée d'étude.

Modalités de soumission des propositions de communication (format 20 minutes + 10 minutes de questions) :

Les propositions de communication devront être rédigées en Times New Roman, interligne simple, taille 12, faire 1 page maximum (bibliographie sélective comprise) et comporter 5 mots-clés. Les propositions devront être anonymisées avant envoi. Nous encourageons les participant.es à présenter leur communication en français. Les communications en anglais seront néanmoins prises en compte.

La soumission des propositions se fait directement sur le site langaje2024.sciencesconf.org au format Word et au plus tard le **dimanche 5 mai 2024**. Lors de la journée d'étude, les intervenant.es seront invité.es à participer au volet vulgarisation selon des modalités qui seront précisées ultérieurement.

Dates clés :

• 5 mai 2024 : date limite de soumission des propositions

• Début/mi- juillet 2024 : notification de l'acceptation par le comité scientifique

• Vendredi 15 novembre 2024 : journée d'étude

Le comité d'organisation
Aure Espilondo (CEL)
Caroline Crépin (CEL)
Catline Dzelebdzic (CeRLA)
Marius François (CeRLA)
Emma Giraudier (CeRLA)
Aurélie Héois (CEL)
Ahmed Mahdi (CeRLA)
Melissa Martin-Kemel (CEL)
Lucky Nte (CeRLA)
Marie-Alice Rebours (CeRLA)
Lisa Stepanian (CeRLA)
Olga Tarabanova (CEL)
Iuliia Troitskaia (CEL)
Amandine Vattaire (CEL)
Emilie Vilmen (CeRLA)

Bibliographie sélective :

ABOUDA Lofti & BAUDE Olivier, 2006, « Constituer et exploiter un grand corpus oral : choix et enjeux théoriques, le cas des ESLO 1 », communication présentée au colloque « Corpus en Lettres et Sciences sociales, des documents numériques à l'interprétation » à Albi (https://shs.hal.science/halshs-01162506).

BIBER Douglas, 1993, « Representativeness in Corpus Design », *Literary and Linguistic Computing,* vol. 8, n°4, 243257.

BLANCHE-BENVENISTE Claire, 2007, « Corpus de langue parlée et description grammaticale de la langue », *Langage et société,* vol. 3-4, nº122-123, 129-141.

CONDAMINES Anne, 2005, « Linguistique de corpus et terminologie », *Langages,* n°157, 36-47.

Corpus du Français Parlé de nos Régions (CFPR), équipe EA 4509 STIH (Sens, Texte, Informatique et Histoire) de la Sorbonne Université, à l'initiative de Mathieu Avanzi et André Thibault (https://cfpr.huma-num.fr/).

GRANGER Sylviane & LEFER Marie-Aude, 2020, « The Multilingual Student Translation corpus: a resource for translation teaching and research », *Language Resources and Evaluation,* vol. 54, nº4, 1183-1199.

HASSE Anja, BACHMANN Sandro & GLASER Elvira, 2019, « *Gschmöis* – Crowdsourcing grammatical data of Swiss German », *Linguistics Vanguard*, vol. 7, 1-17.

JOUITTEAU Mélanie, 2012, « La linguistique comme science ouverte », *Lapurdum*, vol. 16. DOI : https://doi.org/10.4000/lapurdum.2357

LÉON Jacqueline, 2008, « Aux sources de la « Corpus Linguistics » : Firth et la London School », *Langages*, vol. 171, n°3, 12-33.

MEUNIER Mariette, 2008, « La constitution d'un corpus, parcours initiatique en linguistique ». (identifiant HAL : halshs- 00359903).

MILLOUR Alice & FORT Karën, 2019, « Sciences participatives et diversité linguistique. Retours d'expériences », *in Recherche culturelle et sciences participatives*, Culture et Recherche (ministère de la Culture), vol. 140, 90-91.

PLOUX Sabine, GENAY Michael & PLOUX-CHILLÈS Leu, 2021, « Les mots du Grand Débat national : les réseaux lexicaux des contributions déposées sur trois plateformes », *Humanités numériques*, vol. 4. DOI : https://doi.org/10.4000/revuehn.2655

Projet Écriture Scolaire et Universitaire : Corpus, Analyses Linguistiques, Modélisations Didactiques (E-CALM), université Sorbonne Nouvelle, université Paris Saint-Denis, université Toulouse Jean-Jaurès, université Grenoble Alpes (https://e-calm.huma-num.fr/).

PURSCHKE Christoph, 2017, « Crowdsourcing the linguistic landscape of a multilingual country. Introducing Lingscape in Luxembourg », in *Linguistik Online*, vol. 85, nº6, 181-202. DOI : https://doi.org/10.13092/lo.85.4086

TREFFORT Cécile, 2014, « Le corpus du chercheur, une quête de l'impossible ? Quelques considérations introductives », in *Le corpus. Son contour, ses limites et sa cohérence*, *Annales de Janua*, nº2 (http://annalesdejanua.edel.univ-poitiers.fr/index.php?id=725).