Langa[-]

Linguistique, traductologie, didactique des langues

sciences ouvertes

Données et corpus en linguistique à la lumière de la science ouverte : problématiques et enjeux méthodologiques

Vendredi 15 novembre 2024

Salle Marc Bloch (MSH)
14 Avenue Berthelot, 69007 Lyon

Comité d'organisation

Caroline Crépin, CEL, Université Jean Moulin Lyon 3
Catline Dzelebdzic, CeRLA, Université Lumière Lyon 2
Aure Espilondo, CEL, Université Jean Moulin Lyon 3
Marius François, CeRLA, Université Lumière Lyon 2
Emma Giraudier, CeRLA, Université Lumière Lyon 2
Aurélie Héois, CEL, Université Jean Moulin Lyon 3
Ahmed Mahdi, CeRLA, Université Lumière Lyon 2
Melissa Martin-Kemel, CEL, Université Jean Moulin Lyon 3
Lucky Nte, CeRLA, Université Lumière Lyon 2
Marie-Alice Rebours, CeRLA, Université Lumière Lyon 2
Lisa Stepanian, CeRLA, Université Lumière Lyon 2
Olga Tarabanova, CEL, Université Jean Moulin Lyon 3
Iuliia Troitskaia, CEL, Université Jean Moulin Lyon 3
Amandine Vattaire, CEL, Université Jean Moulin Lyon 3
Emilie Vilmen, CeRLA, Université Lumière Lyon 2

Enseignant·e·s référent·e·s

Séverine Wozniak, CeRLA, Université Lumière Lyon 2
Denis Jamet, CEL, Université Jean Moulin Lyon 3

Programme de la journée

9.00	Accueil café
	Conférence plénière
9.30-10.30	**Christophe Benzitoun**, Université de Lorraine et ATILF, Julie Glikman, Université de Lorraine et ATILF, Nicolas Mazziotta, Université de Liège et UR Traverses, Camille Fauth, Université de Strasbourg et LiLPa, « Les vocaux : de la constitution à l'exploitation d'un corpus de pratiques langagières émergentes »
10.30-11.00	Pause café
	Session 1
11.00-11.30	**Gabriella de Luca**, Université Paris-Est Créteil, « Collecte des données dans les archives universitaires pour constitution de corpus à l'égard de la science ouverte : questions éthiques et juridiques »
11.30-12.00	**Evgenia Nicol-Bakaldina**, Université des Antilles, « *Sound of silence* ou les données qui « parlent » : gestion du bruit dans un corpus multimodal des apprenants »
12.00-13.30	Déjeuner
	Session Poster
13.30-14.15	**Manon Anzeraey,** Université Paris Nanterre et Université Paris-Est Créteil, « Constitution et exploitation d'un corpus d'apprenants francophones à partir de productions écrites authentiques d'anglais L2 »

Lucas Maison, Université d'Avignon, « Comment faire pour améliorer la collecte de données audio ? Une analyse des biais du corpus CommonVoice »

Ali Wafdi, Université de Lorraine, « Construction d'un corpus multimodal dans le cadre d'une ethnographie multisituée du dispositif ELCO dans l'Hérault »

Session 2

14.15-14.45	**Lisa Stepanian**, Université Lumière Lyon 2, « Corpus audiovisuels : enjeux et avantages »
14.45-15.15	**Tiago Joseph** et Catherine Bouko, UGent – Université de Gand, « Traiter un corpus féministe et queer issu d'Instagram. Enjeux éthiques, pratiques en jeu »
15.15-15.45	**Christelle Gérard**, Université Paris 8, « Défis éthiques et méthodologiques pour l'élaboration d'un corpus vidéo de langue des signes française impliquant des adultes avec trouble du neurodéveloppement »
15.45-16.00	Pause café

Conférence de clôture

16.00-17.00	**Isabel Colón de Carvajal**, École normale supérieure de Lyon, ICAR, Association Française de Linguistique Appliquée, « Enjeux méthodologiques et juridiques appliqués aux terrains en santé : ne pas rester sur ses acquis ! »
17.00	Remerciements

Conférence plénière

« Les vocaux : de la constitution à l'exploitation d'un corpus de pratiques langagières émergentes »

Christophe Benzitoun, Université de Lorraine et ATILF
Julie Glikman, Université de Lorraine et ATILF
Nicolas Mazziotta, Université de Liège et UR Traverses
Camille Fauth, Université de Strasbourg et LiLPa

Ces dernières années, l'envoi de messages vocaux s'est largement diffusé dans différentes couches de la population. Cette nouvelle pratique est intéressante pour les linguistes car il s'agit de nouvelles données écologiques. Elles permettent donc potentiellement d'observer des phénomènes linguistiques émergents, peu présents dans les entretiens et les conversations présents dans les corpus de français parlé « classiques ».

Dès le départ, l'objectif du projet était de constituer un corpus distribué librement dans différents formats et avec différentes couches d'annotation, en mettant en avant les dimensions phonétique et syntaxique. En tant que support enregistré, le recueil des vocaux peut se faire par simple transfert de l'utilisateur vers le chercheur mais en tant que donnée personnelle, il faut s'assurer de respecter le Règlement Général sur la Protection des Données. Le RGPD nécessite de recueillir en amont le consentement éclairé des participants. Cela a eu une incidence tant sur la procédure que sur le nombre de vocaux collectés et une première phase d'anonymisation et de vérification du contenu des messages a été nécessaire.

Nous détaillerons la chaine de traitement allant de la collecte à l'exploitation, privilégiant au maximum les logiciels libres, tant pour la segmentation que pour l'annotation et l'exploitation. Nous présenterons également quelques exemples d'exploitations possibles. Une première version du corpus est d'ores et déjà disponible sur Ortolang à l'adresse : www.ortolang.fr/market/corpora/lesvocaux.

Session 1

« Collecte des données dans les archives universitaires pour constitution de corpus à l'égard de la science ouverte : questions juridiques »

Gabriella de Luca, Université Paris-Est Créteil

Mots clés : *copie d'examen ; didactique de l'écrit ; archives universitaires ; littéracie universitaire ; science ouverte*

La constitution de corpus d'écrits en contexte d'enseignement (scolaire ou universitaire) se fait généralement auprès des enseignants. Cette pratique rend plus facile les démarches éthiques et juridiques nécessaires pour le partage de corpus en libre accès, telles que la demande d'autorisation pour collecter et exploiter des données personnelles dans une recherche – étape devenue essentielle avec l'entrée en vigueur du Règlement général sur la protection des données (RGPD, 2016) – ainsi que l'autorisation de l'auteur pour publier des documents protégés par la propriété intellectuelle. Dans cette communication, nous présenterons les enjeux de la collecte de données dans les archives universitaires ayant comme objectif la constitution et la diffusion d'un corpus de copies d'examen produites par des étudiants de licence. Quel est le cadre juridique pour la collecte, l'exploitation et le partage des données protégés par la loi ?

Un corpus n'est pas une simple donnée brute, il suppose une construction et doit être raisonné, car il dépend « du point de vue qui a présidé à sa constitution » (Rastier, 2004). Dans le cadre de notre thèse, nous avons constitué un corpus de 450 copies d'examen de licence (de L1 à L3) de trois filières : Histoire, Droit et Sciences et Techniques des Activités Physiques et Sportives (STAPS). Ces copies d'examen seront considérées dans leur intégralité pour nos analyses, envisageant le corpus comme un « objet heuristique » (Mayaffre, 2002).

Dans un souci de rendre ce corpus accessible à la communauté scientifique, ainsi que de permettre aux enseignants de l'explorer pour améliorer leurs pratiques d'enseignement ayant pour but de contribuer au développement de la littéracie universitaire (Delcambre et Lahanier-Reuter, 2002 ; Jacques et Rinck, 2017), nous nous sommes confrontées à plusieurs questions éthiques et juridiques puisque notre collecte des données a été effectuée dans les archives de l'Université Paris Nanterre. Les copies d'examen sont des archives publiques non communicables librement à cause des données personnelles qu'elles contiennent et ne peuvent être consultées que par dérogation. Ayant consulté plus de 2 mille copies, il nous était impossible de contacter toutes les personnes concernées par le traitement de données pour demander leur autorisation.

Ainsi, nous montrerons quel est le cadre juridique qui appuie cette recherche, puisque même avec l'anonymisation des données personnelles il reste la question de la propriété intellectuelle, et comment les chercheurs peuvent se baser sur l'ordonnance n° 2021-1518 concernant la fouille de textes et des données pour exploiter des documents ou des données protégés par la loi.

Code de la propriété intellectuelle. Consulté le 5 avril 2024, à l'adresse https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006069414

Delcambre, I., & Lahanier-Reuter, D. (2012), « Littéracies universitaires : Présentation. Pratiques », *Linguistique, littérature, didactique*, vol. 153-154. DOI : https://doi.org/10.4000/pratiques.1905

Jacques, M.-P., & Rinck, F. (2017), « Un 'corpus de littéracie avancée' : Résultat et point de départ », *Corpus,* vol. 16. DOI : https://doi.org/10.4000/corpus.2806

Mayaffre, D. (2002), « Les corpus réflexifs : Entre architextualité et hypertextualité », *Corpus*, vol. 1. DOI : https://doi.org/10.4000/corpus.11

Ordonnance n° 2021-1518 du 24 novembre 2021 - Légifrance. Consulté le 5 avril 2024, à l'adresse https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000044362034

Rastier, F. (2004), « Enjeux épistémologiques de la linguistique de corpus », *Texto*! (http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html)

Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données) (Texte présentant de l'intérêt pour l'EEE). (n.d.). Consulté le 5 avril 2024, à l'adresse http://data.europa.eu/eli/reg/2016/679/oj/fra

« *Sound of silence* ou Les données qui « parlent » : Gestion du bruit dans un corpus multimodal des apprenants »

Evgenia Nicol-Bakaldina, Université des Antilles

Mots clés : *corpus multimodal ; EMILE ; gestion du bruit ; interlangue ; TAL*

Introduction. La recherche en linguistique de corpus a évolué grâce à l'intégration croissante des logiciels de traitement automatique des langues (TAL), permettant une exploration détaillée de l'interlangue dans des contextes éducatifs multilingues (Tutin et al., s.d ; Rohlfing, 2006 ; Selinker, 1972). L'interlangue est définie comme une « langue intermédiaire que l'apprenant constitue à partir de tous les matériaux à sa disposition – verbaux ou non verbaux – issus de la langue de départ ou de la langue-cible » (Tardieu et Quivy, 2002). Dans les systèmes de traitement de langue, le chercheur est confronté au « bruit », terme désignant tout contenu non standard (Al Sharou et al., 2021), qui peut affecter la transcription et l'analyse des données. Le contenu « non normé » désigne les écarts par rapport à la norme de la langue (Tutin *et al*., s.d).

Problématique. Notre contribution s'inscrit dans la gestion du bruit dans les corpus multimodaux, un élément crucial pour la fiabilité des analyses linguistiques et pédagogiques. Comment constituer et traiter un corpus multimodal (écrit et oral) d'apprenants, étant donné la nature très hétérogène des sources, surtout lorsque les productions des apprenants sont non normées ? Comment établir un compromis raisonnable entre le bruit « utile », susceptible d'enrichir les données, et le bruit « inutile », qui risque de les appauvrir ? Plus important encore, comment distinguer ces deux types de bruit ?

Contexte. Cette contribution s'inscrit dans une recherche doctorale menée à l'Université Savoie Mont-Blanc (Nicol-Bakaldina, 2023). Une classe de section européenne au lycée de Chambéry a été observée pendant une année scolaire, avec des cours d'histoire-géographie en

anglais enregistrés et transcrits, formant le corpus CORINE-SEHEA[1] de 119 000 mots. Cette recherche examine comment, en contexte EMILE (Enseignement d'une Matière par l'Intermédiaire d'une Langue Étrangère), la langue, à la fois outil et objet d'apprentissage, influence les résultats éducatifs.

Méthodologie. La collecte et le traitement des données ont impliqué 16 heures de cours enregistrées et 230 documents divers. Les outils TAL utilisés (Exmaralda, CLAN, AntConc, SketchEngine) ont permis une analyse quantitative et qualitative du corpus, en se concentrant sur la gestion du bruit dans un corpus multimodal d'apprenants en contexte EMILE.

Résultats. Les choix de transcription affectent significativement la qualité des données (Granger et al., 2015). Cette étude utilise des conventions de transcription standardisées (CHILDES ; MacWhinney, 2000) afin d'assurer la cohérence et la comparabilité des données. Les résultats mettent en lumière comment le bruit, lorsqu'il est correctement géré, enrichit l'analyse linguistique sans compromettre la qualité du corpus. L'étude des erreurs des apprenants et de leurs productions non normées révèle des insights sur le fonctionnement de l'interlangue et les processus d'apprentissage.

Conclusion. La recherche souligne l'importance de la rigueur dans la collecte et le traitement des données pour obtenir des corpus représentatifs et fiables. La gestion efficace du bruit est essentielle pour préserver l'intégrité et la pertinence des analyses linguistiques, en adéquation avec les objectifs de la science ouverte et la didactique des langues (Benveniste, 1966 ; Rastier, 2004).

Al Sharou, K., Li, Z., & Specia, L. (2021), « Towards a better understanding of noise in natural language processing », *in Proceedings of the International*

[1] **COR**pus des **IN**teractions à l'**E**cole : **S**ection **E**uropéenne **H**istoire **E**nseignée en **A**nglais

Conference on Recent Advances in Natural Language Processing (RANLP 2021), p. 53-62. https://aclanthology.org/2021.ranlp-1.7

Benveniste, É. (1966), *Problèmes de linguistique générale*, Paris : Gallimard.

Granger, S., Meunier, F., & Gilquin, G. (dir.). (2015,. *Cambridge handbook of learner corpus research*, Cambdrige : Cambridge University Press.

Nicol-Bakaldina, E. (2023), *L'enseignement d'une matière par intégration d'une langue étrangère (E.M.I.L.E) en France : Le rôle et l'utilisation de la langue à l'intersection entre deux disciplines dans l'enseignement secondaire*. Thèse de doctorat, Université Savoie Mont Blanc, LLSETI. https://theses.fr/2023CHAMA028

MacWhinney, B. (2000), *The CHILDES project: Tools for analyzing talk – Electronic edition Volume 1: Transcription format and programs. Part 1: The CHAT transcription format*, Lawrence Erlbaum Associates.

Selinker, L. (1972), « Interlanguage », *International Review of Applied Linguistics*, vol. 10, nº3, 219-231.

Tutin, A., Jaques, M.-P., Kraif, O., & Hartwell, L. (s.d.), *Introduction à la linguistique de corpus*, Université Grenoble Rhône-Alpes. [Cours en ligne de l'UGA].

Session poster

« Constitution et exploitation d'un corpus d'apprenants francophones à partir de productions écrites authentiques d'anglais L2 »

Manon Anzeraey, Université Paris Nanterre et Université Paris-Est Créteil

Mots clés : *linguistique contrastive ; didactique intégrée des langues ; recherche collaborative ; corpus d'apprenants ; enseignement-apprentissage de l'anglais L2 dans le secondaire en France*

Le recueil de données authentiques d'apprenants en anglais L2 en France est soumis aux contraintes intrinsèques du cadre institutionnel de l'enseignement des langues vivantes dans le secondaire et du terrain dans lequel se fait le recueil. La communication proposée ici s'inscrit dans un travail de thèse et porte sur la problématique suivante : dans quelle mesure le recueil de productions écrites d'apprenants en anglais L2 est possible et sous quelles conditions ? Nous souhaitions récolter des productions allant de la 6ème à la terminale afin d'obtenir un relevé représentatif de la réalité du terrain. Nous voulions également que ce recueil se fasse dans des conditions réelles d'apprentissage. Enfin, nous ambitionnions à obtenir un relevé quantitatif d'occurrences des auxiliaires HAVE, BE, DO pour pouvoir en réaliser une analyse qualitative.

Ainsi, notre cadre théorique repose sur la linguistique contrastive et la TOPE (Culioli, 1990 ; 1999) et sur la didactique intégrée des langues. Ces concepts sont selon nous compatibles avec l'approche actionnelle et avec le CECRL (2001). S'est ensuite posée la question méthodologique de recueil des données et de constitution du corpus. En effet, « chaque corpus est motivé par rapport à une fonction spécifique » (Brunet, 2020). En outre, le chercheur doit se prémunir contre une subjectivité

possible afin que les résultats obtenus soient des observables valides (Rabatel, 2013). Des difficultés sont alors apparues.

Les problématiques liées à la recherche collaborative : accords des enseignants et de la hiérarchie, autorisations RGPD, contraintes institutionnelles, faisabilité, nous ont incitées à réduire l'étendue de la collecte à des données écrites anonymisées de lycée (2de à la terminale). Nous avons constitué un premier corpus à partir de ces données. Le sujet d'écriture créative épistolaire a été créé par nos soins. Puis, l'enseignante l'a donné à ses élèves en classe. Le but était de récolter des productions de 100 mots maximum en 30 min. Nous voulions à partir de cette évaluation diagnostique évaluer les acquis et besoins des apprenants afin d'en caractériser la langue. Nous avons récolté 130 productions (13 829 mots), puis restreint notre analyse à la classe de 2de soit 24 productions (2610 mots). Nous avons ensuite analysé qualitativement et quantitativement, au sein du corpus, la réalisation des auxiliaires HAVE, BE, DO par les apprenants.

Brunet, A. (2020), « Quel corpus pour l'identification des compétences des apprenants de niveau intermédiaire et avancé ? Les cas de la cohérence et de la cohésion », *Lidil*, vol.61. DOI : https://doi.org/10.4000/lidil.7424

Conseil de l'Europe. (2001), *Cadre européen commun de référence pour les langues : Apprendre, enseigner, évaluer*. Paris : Éditions Didier.

Conseil de l'Europe. (2012), *Cadre de référence pour les approches plurielles des langues et des cultures : Compétences et ressources*. ISBN : 978-92-871-7172-6.

Culioli, A. (1990 ; 1999), *Pour une linguistique de l'énonciation* (Tomes 1 à 3), Gap : Éditions Ophrys.

Doquet, C., & Ponton, C. (2021), « Écrire de l'école à l'université : Corpus, traitements, analyses outillées. Présentation », *Langue Française*, vol. 211, nº3, 11-20.

Mair, C. (2018), « Contrastive analysis in linguistics », *Oxford Bibliographies*.

Rabatel, A. (2013), « L'engagement du chercheur, entre·"éthique d'objectivité" et "éthique de subjectivité" », *Argumentation et analyse du discours*, vol. 11. DOI : https://doi.org/10.4000/aad.1526

« Comment faire pour améliorer la collecte de données audio ? Une analyse des biais du corpus CommonVoice »

Lucas Maison, Université d'Avignon

Mots clés : *reconnaissance vocale ; Commonvoice ; biais ; réseaux de neurones ; corpus*

Dans le domaine du traitement automatique des langues, il est devenu crucial de disposer de larges quantités de données annotées et de qualité afin d'entraîner les modèles d'apprentissage automatique aux capacités toujours plus grandes.

CommonVoice [1] est un grand corpus de données audio multilingue soutenu par la fondation Mozilla. C'est une initiative open-source de science participative : des volontaires du monde entier peuvent donner leur voix en lisant des phrases dans la langue de leur choix. D'autres volontaires sont alors invités à contrôler les phrases lues pour s'assurer qu'elles correspondent aux textes.

A ce jour, plusieurs centaines de milliers de personnes ont donné leur voix dans plus de 120 langues, totalisant plusieurs dizaines de milliers d'heures d'audio annotées, ce qui représente une ressource précieuse pour les chercheurs en linguistique ou en intelligence artificielle. Des campagnes de collecte de données sont régulièrement organisées par les différentes communautés linguistiques afin de recruter de nouveaux bénévoles et élargir la taille du corpus.

Au cours de nos recherches, nous nous sommes intéressés à la répartition démographique des données collectées, et à la manière dont celle-ci varie entre les différentes langues. Plus spécifiquement, nous avons analysé la distribution des données selon trois attributs démographiques : le genre du locuteur, son âge, et sa région (voix accentuées).

De nombreux travaux ont montré que les modèles de reconnaissance vocale étaient biaisés envers certains groupes : [2] ont montré que le

taux d'erreur mot était supérieur pour les voix féminines, et [3] ont mis en évidence un biais racial. L'étude et la compréhension de ces disparités est un important enjeu de la recherche sur les technologies vocales, l'objectif étant le développement de systèmes qui soient à la fois performants et non-biaisés, c'est à dire fonctionnant avec la même qualité pour tous.

Au travers d'expériences d'apprentissage automatique, nous montrons que les biais de ces modèles semblent majoritairement issus des disparités démographiques des données utilisées pour les entraîner [4]. Une piste prometteuse de réduction de ces biais consiste donc à améliorer la diversité et l'hétérogénéité des corpus. Nous formulons plusieurs recommandations concrètes concernant la meilleure manière d'élargir ou de créer de nouveaux corpus.

[1] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., ... & Weber, G. (2019). Common voice: A massively multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

[2] Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., ... & Goel, S. (2020), « Racial disparities in automated speech recognition », *Proceedings of the National Academy of Sciences*, vol. 117, nº14, 7684-7689. https://doi.org/10.1073/pnas.1914616117

[3] Tatman, R. (2017), « Gender and dialect bias in YouTube's automatic captions », *in* D. Hovy *et al.* (éds.), *Proceedings of the first ACL workshop on ethics in natural language processing*, Valencia : Association for Computational Linguistics, 53-59 DOI : https://doi.org/10.18653/v1/W17-3007

[4] Ngueajio, M. K., & Washington, G. (2022), « Hey ASR system! Why aren't you more inclusive? », *in International Conference on Human-Computer Interaction*, Springer Nature Switzerland, 421-440. DOI : https://doi.org/10.1007/978-3-031-07244-0_34

« Construction d'un corpus multimodal dans le cadre d'une ethnographie multisituée du dispositif ELCO dans l'Hérault »

Ali Wafdi, Université de Lorraine

Mots clés : *dispositif ELCO ; recherche mixte ; linguistique de corpus*

À travers une ethnographie multisituée de l'Enseignement des Langues et Cultures d'Origine (ELCO), nous avons cherché à éclairer ce dispositif éducatif dans le contexte des politiques linguistiques en France. L'évolution que connaît actuellement le dispositif ELCO arabe marocain, l'actuel débat sur la refondation de l'Ecole française, sur l'ancrage des valeurs laïques dans la société de la République, et sur la place du plurilinguisme et du multiculturalisme. Une recherche scientifique appliquée au dispositif ELCO nous a invité à interpeller l'histoire de la genèse de ce projet, en prenant appui sur des sources documentaires que nous avons inventoriées et classifiées. Cette recherche a questionné aussi les acteurs impliqués dans ce dispositif. Ces derniers englobent plusieurs parties prenantes en France et au Maroc. Il s'agit des acteurs politiques, des acteurs opérationnels (inspecteurs, conseillers pédagogiques, directeurs des écoles, enseignants), ceux de la périphérie (élus), les acteurs associatifs, ainsi que les parents des élèves. Nous avons construit notre terrain de recherche auprès de ces acteurs afin de mettre en place des moyens de recherche qui visent à collecter des informations au sein des acteurs impliqués directement ou indirectement dans le cadre de ce projet bilatéral, entre la France et le Maroc.

Pour appréhender le dispositif ELCO, nous avons eu besoin de techniques de recueil de données qui permettent la compréhension des choix et des raisons profondes des acteurs qui structurent ce dispositif. Pour ce faire, nous avons utilisé principalement des techniques qui relèvent de la démarche qualitative, qui sont utilisées en sciences sociales et humaines notamment les entretiens semi-directifs, l'observation participante et l'inventaire des ressources documentaires. Ainsi, nous avons essayé d'explorer, d'analyser et de comprendre, à

travers les témoignages des acteurs, comment s'articulent les décisions entre le bas et le haut dans la mise en place de l'ELCO.

La collecte des données nous a pris trois ans d'engagement sur le terrain, entre bénévolat dans les associations culturelles et cultuelles à Montpellier, et travail d'observation participative au sein des écoles publiques. Entre les établissements scolaires, l'Inspection Académique, les associations culturelles et cultuelles, le Consulat du Maroc à Montpellier et la Fondation Hassan II au Maroc, nous avons tracé le processus de la prise de décision, ce qui nous a permis de recueillir nos données empiriques. Au vu de la taille de notre corpus, nous avons fait le choix d'opter pour une analyse informatisée grâce à un logiciel d'analyse de corpus. Le modèle d'analyse que nous avons mis en place reprend les paramètres du modèle d'analyse de l'action publique de Lascoumes et Le Galès (2012) en plus d'un nouveau paramètre qui semble, de notre point de vue, important d'étudier. Il s'agit de « la symétrie de l'information » au sein du dispositif ELCO.

Plus techniquement, les explorations des observables ont été facilitées par l'usage de la linguistique de corpus (logiciels de traitement de l'oral et de l'écrit).

Arborio, A.-M., & Fournier, P. (2010), *L'enquête et ses méthodes : L'observation directe*, Paris : Armand Colin.

Lascoumes, P., & Le Galès, P. (2012), *Sociologie de l'action publique*, Paris : Armand Colin.

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

Session 2

« Corpus audiovisuels : enjeux et avantages »

Lisa Stepanian, Université Lumière Lyon 2

Mots clés : *traduction audiovisuelle ; localisation ; linguistique ; corpus ; didactique*

La recherche en traduction repose largement sur l'analyse de corpus qui constitue la matière première de l'investigation scientifique. Cependant, la constitution d'un corpus de qualité soulève de nombreux défis, notamment en ce qui concerne l'accès aux données, les droits d'auteur et de diffusion. La préparation de corpus se complique encore plus quand il s'agit de données audiovisuelles car les droits d'image s'y ajoutent, ce qui mène que ces types de corpus sont plus rares (Abouda & Baude, p.2).

Cette proposition de communication vise à illustrer ces défis à travers l'expérience de la constitution d'un corpus de films d'animation traduits spécialement avec des corpus hétérogènes entre la version sous-titrée et les différentes versions doublées.

La constitution du corpus était une démarche complexe et chronophage et elle a nécessité trois étapes majeures :

1. **Sélection et acquisition des données :** La recherche d'une source fiable a conduit à l'utilisation de la plateforme Disney+. Il a fallu faire l'inscription sur la plateforme en France et en Egypte pour accéder aux cinq versions des films sélectionnés : 1) version anglaise, 2) version française et trois versions arabes [3) sous-titrage arabe, 4) doublage arabe standard et 5) doublage égyptien].
2. **Transcription des données :** La transformation du corpus oral en écrit s'est avérée particulièrement laborieuse. L'absence de transcriptions en libre accès, la qualité inégale des transcriptions trouvées en ligne et l'inutilité des logiciels de transcription

automatique face aux environnements bruyants des films d'animation ont nécessité un travail manuel fastidieux.

3. Alignement des données : L'alignement des cinq versions dans un seul tableau a combiné des approches automatiques (Web Align Toolkit) et manuelles (révision).

Le corpus ainsi constitué poursuit deux objectifs principaux :

1. Analyse de la traduction : L'étude portera sur les aspects linguistiques, culturels et techniques de la traduction et de la localisation dans les différentes versions des films, en mettant l'accent sur les choix lexicaux, stylistiques et culturels opérés par les traducteurs.
2. Mise à disposition du corpus en libre accès : Le corpus sera mis en libre accès pour permettre à d'autres chercheurs de l'exploiter dans le cadre de leurs travaux ultérieurs. Il fera également le noyau des cours de didactique de localisation à l'Université d'Alexandrie en présentant une référence pratique, concrète et appliquée pour les jeunes localisateurs ainsi qu'un tutoriel efficace comportant à la fois des règles théoriques et des exemples à l'appui pour faciliter la compréhension du doublage et du sous-titrage (Peromingo, Martín & Riaza), de leur rôle dans la localisation et de leur impact sur le public.

L'expérience de la constitution de ce corpus met en lumière les défis liés à l'accessibilité et à la compilation des données dans le domaine de la linguistique et de la didactique des langues. En partageant notre expérience et en soulevant ces problématiques, nous souhaitons contribuer à la réflexion sur les enjeux méthodologiques et théoriques de la science ouverte dans ces disciplines.

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

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

Peromingo, J. P. R., Martín, R. A., & Riaza, B. G. (2014), « New approaches to audiovisual translation: The usefulness of corpus-based studies for the teaching of dubbing and subtitling », *in* E. Bárcena, T. Read & J. Arús (éds.), *Languages for specific purposes in the digital area*, Berlin : Springer, 145-156. DOI : https://doi.org/10.1007/978-3-319-02222-2_14

Díaz Cintas, J. (2008), *The didactics of audiovisual translation*, Amsterdam : John Benjamins Publishing Company.

« Traiter un corpus féministe et queer issu d'Instagram. Enjeux éthiques, pratiques en jeu »

Tiago Joseph et Catherine Bouko, UGent – Université de Gand

Mots clés : *approche quantitative ; numérique ; données sensibles ; visualisation et annotation ; analyse du discours*

Cette communication constitue un retour transversal sur un cycle complet (de la conception à la première publication) d'une recherche quantitative sur des données numériques, personnelles et sensibles, pour en questionner les enjeux relatifs à la science ouverte. Il s'agit de présenter les défis, les réflexions et les outils techniques accumulés au cours de la constitution d'un corpus de thèse sur les discours féministes et queers sur Instagram, composé de 1900 comptes et 300 000 postes.

Dès la **rédaction du plan de gestion de données**, nous avons été confrontées à plusieurs demandes paradoxales : le RGPD interdit de partager des données personnelles et sensibles, sans les anonymiser préalablement, et demande d'en collecter le minimum nécessaire ; les invitations à la science ouverte appellent à rendre ces données accessibles à des fins de vérification et de réutilisation ; sans avoir force de loi (franzke et al. 2020), les conditions d'utilisation d'Instagram contraignent à récolter ces données manuellement. Face à ces normes, constituer un corpus quantitatif de données de réseaux sociaux devient un dilemme, autant pour la collecte – impossible manuellement – que pour l'anonymisation – rendue caduque par l'indexation des contenus numériques dans les moteurs de recherche. Il s'agit alors de mobiliser les éthiques et épistémologies numériques (Bouko 2024, franzke et al. 2020, Millette 2023), féministes (Abbou & Burnett 2023, franzke 2020) et de linguistique dite folk (Marignier 2019) pour repenser les paradigmes du consentement, de la recherche collaborative, de la distinction privé/publique, de la hiérarchisation des voix, du trio visibilité/responsabilité/propriété de la parole, de la représentativité,

de la pérennité et de la transmission, lorsque ces paradigmes s'appliquent à des données numériques, quantitatives et sensibles.

Au-delà du **recensement de la population** qui soulève autant de questions théoriques (comment définir un compte féministe ?) que méthodologiques (comment cartographier ces comptes dans l'environnement numérique ?), les outils de **collecte** contraignent fortement les terrains et les questionnements possibles. Leur utilisation nécessite des compétences techniques, indispensables pour évaluer la fiabilité de la collecte des données qui requièrent également des outils adaptés à la multimodalité du web. À l'issue de notre recherche, nous souhaitons présenter nos suggestions, nos questions et **DIAMS, notre outil de visualisation et d'annotation de données multimodales**. Nous souhaitons ainsi participer au dialogue collaboratif luttant contre le désarroi et l'éparpillement des ressources, dans une optique de science ouverte dépassant le cadre des laboratoires et des institutions de recherche.

Enfin, la **communication scientifique** constitue parfois un moment paradoxal. Il nous semble important de discuter de l'inconfort, des contraintes et des stratégies d'adaptation que peuvent générer certaines trajectoires de recherche. Face à l'impossibilité d'anonymiser ou d'obtenir un consentement pour des données numériques, quantitatives et sensibles, nous reviendrons sur la possibilité d'utiliser des exemples fictifs, discutée au regard de l'analyse du discours et des études de genre.

Abbou, J., & Burnett, H. (2023), « Des corpus féministes face à l'institution éthique », communication présentée au 3e Congrès International de l'Institut du Genre (université Toulouse Jean Jaurès).

Bouko, C. (2024). *Visual citizenship: Communicating political opinions and emotions on social media*. Routledge.

Franzke, A. S. (2020), « Feminist research ethics », *in Association of Internet Researchers: IRE 3.0 companion 6.3*, 64-75.

Franzke, A. S., Bechmann, A., Zimmer, M., Ess, C., & the Association of Internet Researchers (2020), *Internet research: Ethical guidelines 3.0*.

Marignier, N. (2019), « Les savoirs sur les pratiques langagières féministes et LGBTQI entre académie et militantisme », *Cahiers de l'ILSL*, vol. 28, 87-107.

Millette, M. (2023), « Théorie et méthodologie féministes pour la recherche en contexte numérique », *Communication*, vol. 40, nº, 1-22.

« Défis éthiques et méthodologiques pour l'élaboration d'un corpus vidéo de langue des signes française impliquant des adultes avec trouble du neurodéveloppement »

Christelle Gérard, Université Paris 8

Mots clés : *langue des signes française ; trouble du neurodéveloppement ; trouble du spectre de l'autisme ; corpus-vidéo ; étude de cas multiple*

Dans le cadre d'une étude doctorale interrogeant le rôle de la langue des signes française (LSF) pour la communication de personnes avec un trouble du neurodéveloppement (TND) et considérées comme non verbales, nous avons constitué un corpus vidéo de ± 25 heures. Ceci a impliqué la mise en place d'un atelier d'exposition à la LSF, deux fois par semaine, sur une durée de six mois, pour trois adultes avec TND et leur entourage proche (formatrice de LSF, équipe éducative), atelier que nous avons enregistré en vidéo. Notre objectif étant d'adapter l'approche didactique de cette langue à des personnes avec un TND en nous appuyant sur des situations écologiques de communication, l'enjeu a été d'accéder à des données authentiques et représentatives de l'utilisation de la LSF par ce public spécifique.

Notre recherche, par son aspect pionnier, a suscité des interrogations sur les aspects éthiques à prendre en considération pour l'élaboration et la diffusion de notre corpus impliquant des personnes dites vulnérables. Nous présenterons les démarches entreprises auprès du Comité d'Éthique de notre université et, par la suite, auprès du Conseil de Protection des Personnes.

Puis nous exposerons les problématiques méthodologiques inhérentes à la réalisation d'un tel corpus. Nous avons opté pour une méthodologie empirico-inductive de type qualitatif, l'étude de cas multiple, que nous définirons (Blanchet et Chardenet, 2011), et non-expérimentale (voir entre autres Corbières et Larivière, 2020). Nous avons par ailleurs réduit le paradoxe de l'enquêteur tel que défini par Blanchet (2012) en

adoptant une posture d'observatrice-participante qui nous a permis d'observer à « *l'intérieur de l'interaction langagière et/ou de la communauté linguistique étudiées et donc d'observer des phénomènes habituellement cachés aux 'étrangers'* » (Blanchet, 2012). En effet, dans l'étude de cas, « le chercheur doit savoir se faire accepter dans chacun des sites étudiés » (Gagnon, 2012, p. 58) pour s'assurer de collecter des données riches et crédibles. Nous préciserons ce dernier point.

Finalement, c'est à travers ces questions éthiques et méthodologiques que nous discuterons de la possibilité de mettre en œuvre, dans le cadre de la science ouverte, notre souhait de rendre ce corpus disponible pour d'autres recherches.

Blanchet, P. & Chardenet, P. (2011), *Guide pour la recherche en didactique des langues et des cultures : Approches contextualisées*, Paris : Éditions des archives contemporaines.

Blanchet, P. (2012), *La linguistique de terrain, méthode et théorie : Une approche ethnosociolinguistique de la complexité*, Rennes : Presses Universitaires de Rennes.

Corbières, M., & Larivière, N. (2020), *Méthodes qualitatives, quantitatives et mixtes dans la recherche en sciences humaines, sociales et de la santé* (2e éd.). Québec : Presses de l'Université du Québec.

Gagnon, Y. (2012), *L'étude de cas comme méthode de recherche* (2e éd.), Québec : Presses de l'Université du Québec.

Conférence de clôture

« Enjeux méthodologiques et juridiques appliqués aux terrains en santé : ne pas rester sur ses acquis ! »

Isabel Colón de Carvajal, ENS de Lyon, ICAR, Association Française de Linguistique Appliquée

Après une présentation de l'AFLA, nous proposons un retour d'expériences afin d'aborder les enjeux méthodologiques et juridiques liés aux recueils de données appliqués au domaine de la santé. Même si l'on travaille depuis plusieurs années dans ce champ, il est important de ne pas rester sur les connaissances acquises par le passé. Il faut au contraire actualiser régulièrement ses pratiques, notamment en matière de procédures juridiques afin de ne prendre de retard dans le déroulement d'un projet. Notre retour d'expériences abordera ainsi des enjeux plus larges tels que ceux liés à la science ouverte ou à l'accessibilité des données qui n'est pas toujours évidente avec des données sensibles en santé.

Partenaires

Cette journée d'étude a été financée dans le cadre d'un projet « Junior » du service général de la recherche de l'Université Jean Moulin Lyon 3 que nous remercions. Son organisation a été rendue possible à travers la participation financière, logistique et scientifique de nos partenaires : le Centre d'Études Linguistiques – Corpus, Discours et Sociétés (CEL), le Centre de recherche en linguistique appliquée (CeRLA), l'Université Lumière Lyon 2 et l'Université Jean Moulin Lyon 3.

Le volet vulgarisation de la journée d'étude est le fruit d'une collaboration avec la Maison des Sciences de l'Homme Lyon St-Étienne (MSH) et notamment Christian Dury et Justine Chapelon.