# Constitution de corpus écrits d'apprenants en contexte universitaire

Pratiques, enjeux, perspectives acquisitionnelles Journée d'étude - 18 mars 2022

## Résumés des conférences, communications orales et posters

(par ordre alphabétique, premier auteur)

### 1.Conférences

**Gaëtanelle Gilquin, Université Catholique de Louvain, Belgique**

**Recueil, traitement et analyse des corpus d'apprenants : Défis méthodologiques et enjeux**

Les corpus d'apprenants ont prouvé leur utilité dans le contexte universitaire à plus d'un titre, que ce soit pour étudier les caractéristiques de l'apprentissage d'une langue seconde ou étrangère, ou pour aider à améliorer l'enseignement des langues (voir, par exemple, Granger et al. 2015). Étant donné le nombre important de corpus d'apprenants disponibles à l'heure actuelle, il est possible pour de nombreux enseignants-chercheurs de sélectionner parmi ces corpus celui qui permettra d'atteindre l'objectif poursuivi. Pour d'autres, par contre, le corpus souhaité n'existe pas encore, par exemple parce qu'il concerne une langue peu étudiée dans la recherche sur corpus d'apprenants ou parce que la variété recherchée est très spécifique (ex. production écrite d'apprenants de l'anglais en pharmacologie), voire locale (ex. un enseignant qui souhaiterait avoir accès à un corpus de travaux de ses étudiants de Master). Dans ce cas, l'enseignant-chercheur devra constituer lui-même son corpus d'apprenants, en tenant compte d'un certain nombre de défis méthodologiques. Ce sont ces défis, ainsi que les enjeux de la constitution des corpus écrits d'apprenants en contexte universitaire, qui seront au centre de cette présentation.

En m'inspirant notamment des corpus d'apprenants qui ont été constitués au *Centre for English Corpus Linguistics* de l'Université catholique de Louvain, je passerai en revue les différentes étapes qui mènent à l'exploitation d'un corpus d'apprenants : le recueil des données, d'abord, avec des questions comme la représentativité du corpus ou l'inclusion de métadonnées ; le traitement des données, ensuite, qui couvre entre autres la numérisation (là où c'est nécessaire) et les annotations (bien souvent avec des outils créés pour des variétés natives de la langue) ; et enfin, l'analyse proprement dite, avec par exemple la prise en compte des multiples facteurs pouvant influencer la langue d'apprenants ou le choix d'une norme à laquelle comparer la production langagière d'étudiants. Les défis qui peuvent être rencontrés à chacune de ces étapes seront abordés, notamment dans une perspective pédagogique. De nouveaux types de corpus d'apprenants pouvant apporter un regard différent sur la langue d'apprenants seront également présentés (cf. Gilquin 2021).

Gilquin, G. 2021. Hic sunt dracones: Exploring some *terra incognita* in learner corpus research. In A. Čermáková & M. Malá (eds) *Variation in Time and Space: Observing the World through Corpora* (pp. 65-86). Berlin: De Gruyter.

Granger, S., G. Gilquin & F. Meunier (eds). 2015. *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

**Cristóbal Lozano, Université de Grenade, Espagne**

## Key criteria for the design and compilation of a corpus for L2 acquisition research: Showcasing CEDEL2 (a multi-L1 corpus of L2 Spanish)

Learner corpora (LC) are large, systematic databases of authentic language produced naturalistically by learners of a second language (L2) (Callies & Paquot, 2015; Granger et al., 2015; Le Bruyn & Paquot, 2021; Tracy-Ventura & Paquot, 2021). LC have traditionally targeted L2 English, but the increase in L2 Spanish acquisition and teaching over the past decades has triggered the creation of L2 Spanish corpora (Lozano, 2021b).

I will discuss the importance of SLA-informed learner corpus design by showcasing CEDEL2 (version 2): Corpus de Español como L2 (Lozano, 2021a), which follows Sinclair's (2005) 10 corpus-design principles (content selection, representativeness, contrast, structural criteria, annotation, sample size, documentation, balance, topic, homogeneity). These principles have been adapted to collect SLA-relevant variables (Lozano & Mendikoetxea 2013). CEDEL2

also follows the latest LC recommendations in LC design (Tracy-Ventura, Paquot & Myles, 2021).

CEDEL2 is a multi-L1 corpus of L2 Spanish with learners from typologically (un)related languages (English, German, Dutch, Portuguese, Italian, French, Greek, Russian, Arabic, Chinese, and Japanese), coming from all proficiency levels, diverse learning environments (instructed/naturalistic) and different countries. It currently holds language data from 4,399 participants (1,105,936 words) and data collection for a future version is still ongoing. It is mainly a written corpus though there are samples of spoken language (audios & transcriptions) as well. CEDEL2 also contains several native control subcorpora for comparative purposes.

Crucially, all the learner and native subcorpora have been designed following the same principles and criteria so that full Contrastive Interlanguage Analysis (Granger, 2015) can be carried out. Importantly, CEDEL2 contains large amounts of SLA-motivated metadata (i.e., detailed information about the variables belonging to each speaker and each text) that allow to investigate key aspects in SLA, e.g.: L1 (cross-linguistic influence); proficiency level via a placement test (developmental effects); age of onset to L2 Spanish (critical period and age effects); length of exposure to the L2 (exposure effects); length of residence in a Spanish-speaking country (effects of immersion in naturalistic settings); knowledge and proficiency in other foreign languages (other possible cross-linguistic influence); type of task and task conditions (task effects); etc.

We will finally do a quick demonstration of the functionalities of the CEDEL2's latest web-based search engine which, following the latest trends in Open Science, is freely available and downloadable at http://cedel2.learnercorpora.com.

These features of CEDEL2 corpus and its free web interface are ultimately meant to meet the needs of a wide range of users as suggested by Díaz-Negrillo & Thompson (2013): SLA/LCR researchers, natural language processing scientists, language-teaching practitioners, and materials designers.

Callies, M., & Paquot, M. (2015). Learner Corpus Research: An interdisciplinary field on the move. International Journal of Learner Corpus Research, 1(1), 1-6. https://doi.org/10.1075/ijlcr.1.1.00edi

CEDEL2 (Corpus Escrito del Español como L2), version 2: http://cedel2.learnercorpora.com

Díaz-Negrillo, A., & Thompson, P. (2013). Learner corpora: Looking towards the future. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), Automatic Treatment and Analysis of Learner Corpus Data (pp. 9–29). John Benjamins. https://doi.org/10.1075/scl.59.03dia

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. International Journal of Learner Corpus Research,1(1), 7-24. https://doi.org/10.1075/ijlcr.1.1.01gra

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). The Cambridge Handbook of Learner Corpus Research. Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Le Bruyn, B., & Paquot, M. (Eds.). (2021). Learner Corpus Research Meets Second Language Acquisition. Cambridge University Press. https://doi.org/10.1017/9781108674577

Lozano, C. (2021a). CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research. Second Language Research, 02676583211050522. https://doi.org/10.1177/02676583211050522

Lozano, C. (2021b). Corpus textuales de aprendices para investigar sobre la adquisición del español LE/L2. In M. Cruz Piñol (Ed.), E-Research y español LE/L2: Investigar en la era digital (pp. 138-163). Routledge. http://doi.org/10.4324/9780429433528-9

Lozano, C., & Mendikoetxea, A. (2013). Learner corpora and second language acquisition: The design and collection of CEDEL2. En A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), Automatic Treatment and Analysis of Learner Corpus Data. (pp. 65-100). John Benjamins. https://doi.org/10.1075/scl.59.06loz

Sinclair, J. (2005). How to build a corpus. In M. Wynne (Ed.), Developing linguistic corpora: A guide to good practice(pp. 79-83). Oxbow books.

Tracy-Ventura, N., & Paquot, M. (Eds.). (2021). The Routledge Handbook of SLA and Corpora. Routledge. https://doi.org/10.4324/9781351137904

Tracy-Ventura, N., Paquot, M., & Myles, F. (2021). The future of corpora in SLA. In N. Tracy-Ventura & M. Paquot (Eds.), The Routledge Handbook of Second Language Acquisition and Corpora (pp. 409-424). Routledge. https://doi.org/10.4324/9781351137904

## 2.Communications orales

**Edith DURAND, Laboratoire de Recherche sur le Langage – Université Clermont Auvergne, France**
**Paul LOTIN, Laboratoire de Recherche sur le Langage – Université Clermont Auvergne, France**
**Christine BLANCHARD RODRIGUES, Laboratoire de Recherche sur le Langage – Université Clermont Auvergne, France**

### Méthodologie d'analyses des contributions rédactionnelles lors d'une co-écriture en ligne au niveau universitaire

*Mots-clés — Apprentissage collaboratif médiatisé par ordinateur, pédagogie universitaire, interactions*

Nous présentons la méthodologie d'analyses développée dans le projet Acol@d, ayant pour terrain d'étude un enseignement, dispensé en 3ème année de Licence en Sciences du Langage. Dans ce cours dispensé de manière hybride, les apprenant·e·s de langue

maternelle française et en français langue étrangère sont amenés à co-écrire des articles de vulgarisation scientifique., suivant un processus rédactionnel de planification, mise en texte et révision (Flower et Hayes, 1981 : 365-387). Ce projet vise la compréhension des interactions entre co-auteur·e·s et entre tuteur/trice·s et co-auteur·e·s et de leurs impacts sur la qualité de la co-écriture en ligne. Les données concernent les cours dispensés en 2019 et 2020 avec 116 apprenant·e·s répartis en 29 groupes (4 groupes avec apprenant·e·s non-natifs). Les données comprennent les textes co-écrits sur pads avec un historique dynamique mémorisé et le clavardage sur les pads.

Une première analyse quantitative vise à évaluer la balance de participation entre co-auteurs et comment les variables de participation influent sur la qualité du texte final. Comme démontré par Temperman et al (2017) et Olson et al. (2017), il est attendu que l'équilibre de participation favorise un style de travail fluide ainsi que le développement de compétences rédactionnelles et un texte final de qualité. Les logiciels pour cette analyse sont les suivants : un utilitaire développé au sein de l'équipe (PL) pour extraire le nombre et le type de révision, le nombre de caractères écrits par chaque co-auteur·e, la mise en forme de ces progressions dynamiques sur logiciel Excel et le logiciel Jamovi d'analyse statistique. Une deuxième analyse qualitative vise à mesurer l'influence des commentaires des co-auteurs et des tu-teurs/trices sur la production de texte. Comme démontré par Gonthier et al (2018), il est attendu que les commentaires centrés sur la tâche tels que les commentaires de révision entre co-auteur·e·s ou par le tu-teur/-trice, favorisent la progression de textes et participent à la qualité finale du texte. En plus de l'utilitaire développé dans l'analyse 1, le logiciel NVivo est utilisé pour l'analyse qualitative du type de commentaires.

La méthodologie présentée permet une analyse du processus de co-écriture et l'étude des facteurs de réussite ou d'échec dans une tâche de co-écriture, avec une attention particulière portée aux groupes incluant les apprenant·e·s non-natifs. Nos analyses à venir permettront de dégager des recommandations pratiques pour la conduite de projets d'écriture collaborative.

Flower, L., & Hayes, J.R. (1981) A cognitive process theory of writing, *Coll. Compos. Commun., 32* (4)

Temperman, G., Walgraeve, S., de Lièvre, B., & Boumazguida, K. (2017). Développer des compétences de conceptualisation et d'analyse avec un forum de discussion et un etherpad, *Sci. Technol. l'Information la Commun. pour l'Éducation la Form*.

Olson, J. S., Olson, G. M., Zhang, J. & Wang,D. (2017). How People Write Together Now: Beginning the Investigation with Advanced Undergraduates in a Project Course, *ACM Trans. Comput. Interact*, *24*(4).

Gonthier, M. È., Ouellet, C., & Lavoie, N. (2018). Clavardage pédagogique: performances en écriture et interactions entre pairs chez des élèves du secondaire en difficulté d'apprentissage. *Apprentissage des Langues et Systèmes d'Information et de Communication*, 21.

**Thomas GAILLAT, LIDILE – Université de Rennes, France**
**Elisabeth RICHARD , LIDILE – Université de Rennes, France**
**Marie-Françoise BOURVON,  LIDILE – Université de Rennes, France**
**Griselda DROUET, LIDILE – Université de Rennes, France**

## Le Corpus InterLangue, un corpus bilingue comparable

*Mots-clés — Corpus d'apprenant bilingue, français L2, anglais L2 — bilingual learner corpus, French L2, English L2*

L'interlangue des apprenants est le produit de processus cognitifs complexes. Elle peut s'analyser en fonction de trois dimensions que sont l'exactitude, la complexité et la fluence (Housen et al., 2012). Par l'observation et la comparaison de données combinant différentes langues d'apprentissages (L2) et différentes langues natives (L1) (Granger, 1996), il est possible d'identifier des caractéristiques d'usages et schémas erronés communs qui apportent des éclairages sur les différents stades de développement.

Ce type d'approche implique l'utilisation de corpus d'apprenants aux données comparables. Leur comparabilité est un enjeu qui dépend de plusieurs critères tels que les L1 et L2 des locuteurs, les types de tâche, leur durée ou encore les genres. Une solution réside dans l'utilisation de corpus multilingues tels que le corpus MERLIN (Wisniewski et al., 2013), composé de trois modules en allemand, italien et tchèque articulés autour de tâches similaires. Ces textes, qui sont en outre annotés selon le même schéma, offrent des points de comparaisons multiples. Accroître ce type de corpus présente un intérêt évident pour la communauté. La création de corpus d'apprenants oraux multilingues irait dans ce sens.

Fruit de plus d'une décennie de collecte auprès de locuteurs non-natifs, nous présentons le corpus InterLangue (CIL) (Arbach, 2015), dans sa version nettoyée et requêtable (Gaillat et al., 2021) par le biais d'une base de données Huma-Num Nakala. Ce corpus oral et écrit est composé de 115 locuteurs apprenants d'anglais et de français langue étrangère. A chaque apprenant correspondent des fichiers audios retranscrits et des fichiers textes, produits à la suite d'une tâche conversationnelle de vingt minutes environ et d'une tâche narrative de rédaction écrite. L'une des originalités provient des types de tâches qui sont communes aux

deux langues. Grâce à l'existence d'une interface de programmation (API), les différents éléments du corpus peuvent être interrogés et filtrés afin de créer des jeux de données. Des analyses quantitatives peuvent ensuite être entreprises à l'aide de scripts automatisés.

Arbach, N. (2015). *Constitution d'un corpus oral de FLE : enjeux théoriques et méthodologiques* [Phd thesis, Université Rennes 2]. https://tel.archives-ouvertes.fr/tel-01147632

Gaillat, T., Contreras Roa, L., & Attoumbre, J. (2021, septembre). A data repository for the management of dynamic linguistic datasets. *CLARIN Annual Conference 2021*. CLARIN, Madrid (online), Spain. https://hal.archives-ouvertes.fr/hal-03343010

Granger, S. (1996). From CA to CIA and Back : An Integrated Approach to Computerized Bilingual and Learner Corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Éds.), *Languages in Contrast. Text-based cross-linguistic studies* (Vol. 88, p. 37-51). Lund University Press.

Housen, A., Kuiken, F., & Vedder, I. (Éds.). (2012). *Dimensions of L2 performance and proficiency : Complexity, accuracy and fluency in SLA* (Vol. 32). John Benjamins Publishing Company.

**Thomas GAILLAT, LIDILE – Université de Rennes, France**

## Un module MOODLE pour la collecte systématisée de corpus d'apprenants

*Mots-clés — Conception de corpus d'apprenant, metadonnées, structuration, MOODLE*

La constitution d'un corpus d'apprenants est un processus systématique dont l'objectif est de recueillir un échantillon représentatif de la langue d'apprentissage (L2). Ce travail implique la définition de critères externes permettant de mettre en regard les productions recueillies avec les caractéristiques socio-éducatives des sujets observés (Gilquin, 2015). Grâce à ce type de métadonnées, il est alors possible d'échantillonner un corpus en strates, ce qui favorise la représentativité des données à observer (Biber, 1993; McEnery & Hardie, 2012). Ces métadonnées permettent aussi de générer des jeux de données dans lesquels

certaines variables sont contrôlées, et ainsi de favoriser la comparabilité entre cohortes (Burnard, 2005). Par conséquent, la qualité de la représentativité dépend en partie de la qualité des métadonnées retenues.

Dans ce cadre, l'une des difficultés réside dans le choix des critères. Leurs valeurs et portées peuvent varier en fonction des sujets observés, et ce d'autant plus si les informations sont saisies librement par les sujets eux-même. Il n'est donc pas simple de définir un protocole de recueil permettant d'éviter les ambiguïtés concernant les métadonnées collectées. Une autre difficulté du processus de recueil concerne la mise en relation des métadonnées avec les données textuelles. Des données hétérogènes sur des supports distincts rendent difficiles la mise en correspondance ultérieure pour la création d'un jeu de données.

Une solution consiste à mettre en place un dispositif de recueil permettant la création automatique de fichiers reliant métadonnées et données. Cela peut se faire en utilisant des formulaires électroniques de saisie comme pour le corpus COREFL (Lozano et al., 2021). Dans ce cas, les données sont homogènes et classées. Cependant, elles se trouvent sur des serveurs externalisés. Notre proposition est une solution similaire mais permettant la collecte, le stockage et l'accès au sein même des institutions de collecte. Il s'agit d'un module MOODLE (Dougiamas & Taylor, 2003), disponible en ligne en licence Creative Commons et utilisé dans le cadre du projet VizLing (Gaillat et al., 2021). Il offre une interface de saisie dont les métadonnées et les valeurs possibles sont prédéfinies suivant Gilquin (2015). Après saisie, les données et métadonnées L2 sont stockées dans une base pouvant générer tout type de fichier tableur. Elles restent confidentielles et accessibles par les utilisateurs. Ce module présente l'avantage de pouvoir être importé sur toute plateforme MOODLE, et de permettre la collecte de corpus comparables dans différentes institutions.

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, *8*(4), 243-257.

Burnard, L. (2005). Metadata for corpus work. In M. Wynne (Éd.), *Developing Linguistic Corpora : A Guide to Good Practice*. Oxbow.

Dougiamas, M., & Taylor, P. (2003). Moodle : Using Learning Communities to Create an Open Source Course Management System. *Proceedings of the EDMEDIA 2003 Conference, Honolulu, Hawaii*, 171-178. https://www.learntechlib.org/primary/p/13739/

Gaillat, T., Knefati, A., & Lafontaine, A. (2021). Towards a Data Analytics Pipeline for the Visualisation of Complexity Metrics in L2 writings. *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, 123-129. https://www.aclweb.org/anthology/2021.bea-1.13

Gilquin, G. (2015). From design to collection of learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Éds.), *The Cambridge Handbook of Learner Corpus Research*. Cambridge University Press.

Lozano, C., Díaz-Negrillo, A., & Callies, M. (2021). Designing and compiling a learner corpus of written and spoken narratives : COREFL. In C. Bongartz & J. Torregrossa (Éds.), *What's in a Narrative ? Variation in Story-Telling at the Interface between Language and Literacy* (p. 21-46). Peter Lang. https://doi.org/10.3726/978-3-653-05182-7

McEnery, T., & Hardie, A. (2012). *Corpus linguistics : Method, theory and practice*. Cambridge University Press.

**Tomoko HIGASHI, LIDILEM - Université Grenoble Alpes, France**
**Cécile FRÉROT, LIDILEM - Université Grenoble Alpes, France**
**Sylvain COULANGE, LIDILEM - Université Grenoble Alpes, France**
**Hisae AKIHIRO, TUFS, Japon**

## Constitution d'un corpus d'apprenants bilingue et multidimensionnel à partir d'une tâche télécollaborative de traduction

*Mots-clés — télécollaboration, traduction/révision, corpus bilingue et multidimensionnel*

C'est dans le contexte de l'enseignement de la traduction spécialisée, et de l'apprentissage du français que nous œuvrons à la conception d'un cadre d'apprentissage collaboratif entre étudiants français (Université Grenoble Alpes, désormais UGA) et japonais (Tokyo University of Foreign Studies, désormais TUFS). À travers le recueil de données issues d'un travail télécollaboratif, nous explorons les modalités d'apprentissage et d'entre-aide chez les étudiants dans une activité de tandem à distance.

Les échanges oraux et écrits autour d'une tâche collaborative de traduction ont permis la constitution d'un corpus d'apprenants bilingue et multidimensionnel comprenant des discussions en visioconférence portant sur la révision des traductions, en plus des textes traduits et annotés par les étudiants. La tâche collaborative entre étudiants français et japonais, répartis en binômes, porte sur la traduction de textes à dominante socio-culturelle vers leur langue cible, et implique différentes étapes, notamment la révision (Artero et Hamon 2018), prise en charge par l'étudiant locuteur natif. Les interactions réalisées sur Zoom nous offrent ainsi des données intéressantes sur la modalité de négociation lors de la traduction/révision, le processus linguistique et cognitif ainsi que l'interculturalité.

La télécollaboration et le recueil des données s'effectuent sur 3 années consécutives depuis l'année pilote (2019-20).

La transcription et l'annotation des conversations Zoom est en cours. Chaque phrase sur laquelle porte la discussion est renseignée, de manière à pouvoir croiser facilement chaque modification de la traduction avec la négociation orale des étudiants.

Les traductions fournies par les étudiants sont au format Word ou LibreOffice et sont accompagnées de commentaires et du suivi de modifications. Ces documents sont ensuite traités et harmonisés pour aboutir à un document XML plus facilement exploitable.

Une plateforme en développement permet déjà de visualiser l'évolution des traductions dans le temps avec les différentes modifications et annotations des étudiants, et affichera bientôt la transcription de leurs échanges oraux pour chaque phrase. Nous souhaitons ensuite enrichir les annotations : automatiquement, par une analyse morphosyntaxique automatique dans les deux langues ; et manuellement, en permettant à des collaborateurs experts d'étiqueter les annotations en fonction du type de modification.

Nous souhaitons à terme permettre à l'utilisateur d'effectuer des recherches transversales dans l'ensemble du corpus, en fonction de certains types d'erreurs ou de corrections, et, au delà de nos analyses, permettre à des enseignants, des chercheurs ou des étudiants d'utiliser ce corpus, voire d'importer leurs traductions sur la plateforme pour pouvoir visualiser/enrichir les annotations et effectuer des recherches.

Artero, P. & Hamon, Y. (2018). Révisions collaboratives croisées en ligne : apprendre à réviser à plusieurs et à distance, R*evue internationale de pédagogie de l'enseignement supérieur*, 34(2).

Granger, S (Éd) (2015). Learner corpus research, Cambridge University Press.

ZhaoHong H. & Tarone, E. (Éd) (2014). Interlanguage : Forty years later, John Benjamins Publishing Company.

Cutler, C. & Røyneland, U. (2018). Multilingual youth practices computer mediated communication, Cambridge University Press.

**Luca PALLANTI, LIDILEM – Université Grenoble Alpes, France**

## Le corpus d'apprenants : un outil d'évaluation des dispositifs didactiques expérimentaux centrés sur l'écriture ?

*Mots-clés — corpus d'apprenants, écriture, évaluation, didactique*

L'utilisation des corpus d'apprenants pour analyser des dispositifs de formation à l'écriture remonte au début des années 2000 (Chambers & O'Sullivan, 2004). De telles perspectives

ont poussé les chercheurs à utiliser les corpus d'apprenants comme des outils d'évaluation de l'écriture académique (Callies, 2015). Notre étude se propose d'explorer dans quelle mesure les corpus d'apprenants peuvent jouer un rôle dans l'évaluation même des dispositifs didactiques expérimentaux centrés sur les compétences rédactionnelles. Après avoir présenté le corpus ÉNEPS, nous expliciterons les procédures d'annotation et les indicateurs de réussite retenus. Nous exposerons enfin les résultats de l'annotation et comment ils peuvent participer à une évaluation globale de notre dispositif.

Le corpus ÉNEPS, qui compte environ 35.500 mots, a été recueilli dans le cadre d'une expérimentation didactique à l'École Nationale de l'Enseignement Professionnel Supérieur de Grenoble (ÉNEPS), menée avec des étudiants de L1 issus d'un Bac professionnel. Cette étude porte sur le sous-corpus *Synthèses*, qui contient les textes produits par les étudiants lors d'un pré et d'un post-test par un groupe d'étudiants dit *Expérimental* (N=36) et par un groupe d'étudiants dit *Contraste* (N=13) (Boch et al., 2016).

Notre dispositif étant centré sur le développement des compétences de gestion textuelle, nous avons conçu un modèle d'annotation de corpus basé sur les chaines de référence, dont la bonne gestion participe à la fois de la cohésion et de la cohérence textuelles (Charolles, 1995). Les post-tests du groupe *Expérimental* montrent une augmentation significative du nombre moyen de maillons (ou référents) par texte (+6 maillons environ) et du nombre moyen de mots par texte (+37 mots environ). De même, nous avons relevé une diminution significative du nombre moyen de mots entre deux maillons (-4,5 mots entre deux maillons), ce qui signifie que les maillons sont plus rapprochés. À travers des graphiques avec repères de densité, nous avons pu montrer que les textes produits lors des post-tests par le groupe *Expérimental* se caractérisent par une distribution des maillons plus homogène.

Des analyses qualitatives, que nous prendrons le soin de détailler, confirment ces résultats. Les données de notre étude montrent que les corpus d'apprenants peuvent devenir un outil d'évaluation efficace des compétences de gestion des référents (Pallanti et al., 2021). Nous pensons enfin que ces indicateurs pourraient aussi constituer un outil d'évaluation de l'efficacité des dispositifs didactiques visant l'amélioration des compétences de gestion textuelle.

Boch, F., Sorba, J., & Bessonneau, P. (2016). Évaluer les compétences rédactionnelles : Que tester ? *Le français aujourd'hui*, *193*, 127-144. https://doi.org/10.3917/lfa.193.0127

Callies, M. (2015). 2015. "Using learner corpora in language testing and assessment : Current practice and future challenges". In E. Castello, K. Ackerley, & F. Coccetta (Éds.), *Studies in Learner Corpus Linguistics : Research and Applications for Foreign Language Teaching and Assessment*(p. 21-35). Peter Lang.

Chambers, A., & O'Sullivan, Í. (2004). Corpus consultation and advanced learners' writing skills in French. *ReCALL*, *16*(1), 158-172. https://doi.org/10.1017/S0958344004001211

Charolles, M. (1995). Cohésion, cohérence et pertinence du discours. *Travaux de Linguistique: Revue Internationale de Linguistique Française*, De Boeck Université, 125-151.

Pallanti, L., Jacques, M.-P., & Brissaud, C. (2021). Travailler l'écrit pour favoriser la réussite des étudiants issus de bacs professionnels : Un enjeu de linguistique appliquée. *éla. Études de linguistique appliquée*, *2*(202), 167-179.

**Azadeh PIROOZ, LIDILEM - Université Grenoble Alpes, France**

## Évaluation des compétences collocationnelles dans le discours académique : méthode de recueil et d'analyse d'un corpus d'apprenants avancés de français

*Mots-clés — collocations transdisciplinaires, écrit académique, corpus d'apprenants*

À l'université, l'écrit est l'un des outils de communication et d'évaluation des processus d'apprentissage et d'acquisition d'un savoir scientifique. Ce genre de discours obéit à des normes précises et exige des compétences particulières. Les étudiants allophones arrivant en France et issus de systèmes d'enseignement différents avec des compétences linguistiques souvent peu stabilisées, ont des difficultés rédactionnelles quant à l'organisation du texte ou l'association des unités lexicales. Cette dernière difficulté vient en particulier du manque de connaissance des lexicales typiquement utilisées par la communauté scientifique renvoyant aux textes scientifiques eux-mêmes. Il s'agit du lexique et des associations renvoyant aux objets scientifiques, aux procédures d'investigation et au raisonnement relevant de l'activité scientifique. Parmi les compétences lexicales du scripteur, les collocations – qui sont pour nous des cooccurrences privilégiées de deux unités lexicales liées par une relation syntaxique - occupent une place importante dans la langue de spécialité. Il s'agit des expressions « qui renvoient [...] non seulement aux procédures, démarches, objets scientifiques, mais aussi aux éléments d'argumentation, d'évaluation et de structuration du discours » (Jacques & Tutin, 2018 : 6), comme *avancer une hypothèse,* ou *observation empirique.* Ces unités lexicales transdisciplinaires n'appartiennent pas à une discipline précise (il ne s'agit pas de terminologie), mais elles peuvent être utilisées dans les écrits scientifiques (Jacques & Tutin, 2018). La véritable difficulté lexicale des apprenants correspond à ce type de lexique en langue de spécialité (Cavalla, 2008). Cependant, peu de travaux ont porté sur ces phénomènes dans la langue académique des apprenants, en particulier pour le français, alors que ces compétences paraissent centrales.

Nous nous intéressons, ici, aux collocations transdisciplinaires de type V+N (*émettre une hypothèse*) relevées dans 30 productions écrites en français par des mastérisants non natifs. Nous souhaitons, par l'observation et l'analyse des collocations retenues,identifier plus finement les difficultés rencontrées par les étudiants. À cette fin, nous avons élaboré une grille d'analyse des erreurs collocationnelles. Le repérage du type d'erreurs par les évaluateurs et l'analyse est actuellement en cours. Les premiers résultats montrent que les erreurs sémantiques occupent la première place. Une partie importante du désaccord observée au niveau sémantique correspond à l'erreur du type « collocatif inadapté mais sens adapté ».

Cavalla, C. (2008). Les collocations dans les écrits universitaires : un français spécifique pour les apprenants étrangers. https://hal.archives-ouvertes.fr/hal-00397684/document

Jacques, M.-P. et Tutin, A. (2018). Le lexique scientifique transdisciplinaire. Dans M. P. Jacques et A. Tutin (dir.), *Lexique transdisciplinaire et formules discursives des sciences humaines* (p.1-26). ISTE éditions.

## 3.Posters

**Sülün AYKURT-BUCHWALTER, LIDILEM, Université Grenoble-Alpes, France**

### Concevoir et constituer un corpus d'apprenants turcophones du FLE: enjeux méthodologiques

*Mots-clés — corpus d'apprenants, acquisition, écriture en L2*

Les apprenants turcophones du FLE rencontrent des difficultés spécifiques au niveau la production écrite. Une partie de ces difficultés peuvent être liées à l'influence de la L1. Notre recherche a pour but d'élucider les différences et les similitudes entre certaines tendances d'écriture en français et en turc, afin de tenter de comprendre dans quelle mesure les tendances d'écriture et la grammaire de la L1 sont susceptibles d'avoir un impact sur la production écrite en FLE.

Pour cela, nous avons constitué un corpus de scripteurs natifs et apprenants. Le corpus est composé de quatre groupes de scripteurs: les natifs francophones, les natifs turcophones, les turcophones apprenants du FLE au niveau B1, et les apprenants au niveau B2. Pour

chaque groupe de scripteurs, le corpus est constitué de textes appartenant à deux genres différents : des lettres formelles et des courriers électroniques.

Nous proposons, dans ce poster, de décrire les défis méthodologiques liés à la constitution de ce corpus sous trois axes:

- La longitudinalité: Nos groupes d'apprenants B1 et B2 ne sont pas composés des mêmes étudiants suivis à travers leur parcours d'acquisition. De Cock et Tyne (2014) expliquent que dans les études sur l'acquisition, il faut idéalement constituer un corpus longitudinal, mais que cela engendre des défis considérables.
- La représentativité: Des questions se posent quant à la représentativité du corpus. D'une part, les femmes sont surreprésentées dans l'ensemble des groupes. Cela confirme les observations de Gilquin (2015) sur la participation volontaire aux travaux optionnels. D'autre part, nous avons constaté que les groupes B1 et B2 présentaient une certaine hétérogénéité. Il s'agit d'une difficulté méthodologique attestée dans les corpus d'apprenants (Benazzo & Leclercq, 2021).
- La conception des tâches d'écriture: En suivant les recommandations d'Hidden (2014), nous avons souhaité concevoir des sujets appartenant à des genres culturellement familiers à tous les groupes, mais nous n'avons pas échappé à des sensibilités existantes.

Nous expliciterons les solutions que nous avons adoptées face à ces éléments. S'il n'est pas possible de trouver une solution à l'ensemble de ces défis et biais, nous considérons qu'une prise de conscience des enjeux contribue à mitiger leur impact sur les résultats de la recherche. Ce type de réflexion peut contribuer à renforcer les bonnes pratiques pour la constitution d'un corpus écrit dans la lignée des travaux du consortium CORLI, qui propose des recommandations concrètes pour les corpus oraux.

Benazzo, Sandra & Pascale Leclercq (2021). Étudier l'acquisition en L2: Quelles démarches méthodologiques? In Pascale Leclercq, Amanda Edmonds & Elisa Sneed German (Hrsg.), *Introduction à l'acquisition des langues étrangères*. Louvain-la-Neuve: De Boeck Supérieur.

Blanchet, Philippe & Chardenet, Patrick (2017). *Guide pour la recherche en didactique des langues et des cultures: approches contextualisées*. Paris: Editions des archives contemporaines.

CORLI, Bonnes pratiques pour la constitution de corpus https://corli.huma-num.fr/bonnes-pratiques-pour-la-constitution-de-corpus/# consulté le 7 décembre 2021

De Cock, Sylvie & Henry Tyne (2014). Corpus d'apprenants et acquisition des langues. *Recherches en didactique des langues et des cultures* 11(1). 1–23.

Gilquin, Gaëtanelle (2015). From design to collection of learner corpora. In Sylviane Granger, Gaetanelle Gilquin & Fanny Meunier (Hrsg.), *The Cambridge Handbook of Learner Corpus Research*, 9–34. Cambridge: Cambridge University Press.

Hidden, Marie-Odile (2014). *Pratiques d'écriture. Apprendre à rédiger en langue étrangère*. Hachette Français Langue Etrangère.

**Kátia BERNARDON DE OLIVEIRA, Université Grenoble Alpes, France**
**Luciane BOGANIKA, Laboratoire LIDILEM-Université Grenoble Alpes, France**

## L'emploi du verbe ficar dans les manuscrits des étudiants universitaires de PLE en France

*Mots-clés — verbe* ficar*, PLE, Projet CALMER, polysémie.*

Ce poster portera sur le travail qui a été mené jusqu'à présent avec les étudiants de portugais langue étrangère (PLE) issus du secteur LANSAD de l'université Rennes 2 et du Service de Langues de l'université Grenoble Alpes (UGA). Le but de ce travail est de réfléchir sur la dichotomie entre les verbes *ficar* en portugais et *rester* en français dans l'enseignement-apprentissage de PLE. A partir d'un premier échantillon de l'emploi des verbes *ficar* et *rester* par des locuteurs de portugais et de français L1, nous avons vérifié les représentations que les étudiants de PLE ont de ces deux verbes. Notre méthodologie de recherche fait partie du projet CALMER (corpus Comparable pour l'étude de l'Acquisition et les Langues: Multilingue, Émotion, Récit) du laboratoire LIDILEM-UGA. Ce projet vise à créer un corpus multilingue, c'est pourquoi il est ouvert à différentes langues qui souhaitent collaborer à la collecte de données, en suivant, à cet effet, le même protocole. Le dispositif complet de ce projet est disponible sur le site internet d'ORTOLANG (https://www.ortolang.fr/market/corpora/corpus-calmer).

Une première analyse des manuscrits des étudiants de PLE nous a révélé que le verbe *ficar* est utilisé dans différents contextes. De ce fait, le processus d'analyse de l'emploi du verbe *ficar* par les étudiants insérés dans l'espace universitaire français confirme l'importance de poursuivre la recherche sur la dichotomie entre ces deux verbes. En outre, ce projet s'inscrit dans l'analyse du processus d'acquisition linguistique dans une perspective comparative, qui

est en lien avec le projet CALMER. En prenant en compte que les étudiants qui choisissent d'étudier le portugais comme langue étrangère à l'université ont généralement étudié l'espagnol comme langue étrangère (cf. DE OLIVEIRA ; BOGANIKA, 2019), il serait intéressant d'analyser la relation entre les verbes *ficar/rester/quedar*, puisque ce dernier est aussi polysémique et exprime les idées de changement et de permanence (cf. GOMEZ, 2011). Ainsi, dans la continuité de notre projet de recherche, nous souhaitons observer comment les étudiants du PLE ayant le français comme L1, et ayant des connaissances linguistiques en espagnol, s'expriment en portugais L2 : ces trois verbes seraient représentés de manière équivalente ou distincte.

De Oliveira, K. B., & Boganika, L. (2019). La diversité dans l'enseignement/apprentissage du Portugais Langue Etrangère: aspects individuels dans la collectivité. *Reflexos*, (4).

Gómez, L. (2013). L'étrange polysémie du verbe quedar, ou l'expression de la permanence et du changement : étude diachronique et synchronique. *CogniTextes. Revue de l'Association française de linguistique cognitive*, (10).

Gómez, L., & de Oliveira, K. B. (2020). Corpus CALMER: corpus Comparable pour l'étude de l'Acquisition et des Langues: Multilingue, Émotion, Récit.

Perini, M. A. (1995). *Gramática descritiva do português*. Ática.

**Aurélie BOURDAIS, ICAR – Université Lumière – Lyon 2, France**

## Traducteurs en ligne et processus d'écriture en L2 : observation de pratiques invisibles en contexte scolaire

*Mots-clés — traducteurs en ligne – processus – activités cognitives – pratiques buissonnières – captures d'écran dynamiques*

Les outils d'aide à la traduction font aujourd'hui partie de la panoplie numérique des lycéens : ceux-ci y auraient souvent recours lors d'activités de production écrite, dans le cadre d'opérations de mise en texte ou de révision (Hayes & Flower, 1980). Des différences significatives peuvent toutefois être observées dans les discours tenus par les enseignants à propos de ces outils : le dictionnaire en ligne Wordreference serait largement recommandé tandis que les traducteurs en ligne, pourtant massivement consultés par les lycéens, seraient au contraire majoritairement interdits par les enseignants (Bourdais & Guichon, 2020). La consultation de traducteurs apparaît alors comme une pratique clandestine et «

buissonnière », qui donne lieu à des usages « bricoleurs » (De Certeau, 1980), parfois inventifs et plus ou moins pertinents.

L'accès au processus de consultation des différents outils d'aide à la traduction constitue un enjeu majeur qui permettrait d'observer un processus habituellement invisible dans le cadre d'activités de production écrite et de déterminer dans quelle mesure les apprenants ont développé des usages informés de ces outils. L'accès à des pratiques réelles soulève toutefois des questions méthodologiques lorsqu'il s'agit de pratiques clandestines. Le poster conçu pour cette journée d'étude vise à présenter la façon dont la nature clandestine des pratiques étudiées a conduit au recueil de captures d'écran dynamiques selon deux protocoles complémentaires.

La réalisation de captures d'écran dynamiques donne accès à la production écrite en tant que résultat et que processus (Hamel, Séror, & Dion, 2015). Une étude de cas a permis d'analyser les pratiques de deux élèves de Terminale (niveau B2) dans le cadre de trois activités de production écrite. Les élèves pouvaient consulter les outils d'aide à la traduction de leur choix sur leur *smartphone* et ont rédigé le texte en anglais sur papier, afin de respecter l'écologie de leurs pratiques habituelles. Une étude plus expérimentale a ensuite été menée auprès d'apprenants de niveau A2 (n=49) : les élèves ont produit le texte en anglais sur tablette, à l'aide du traducteur en ligne Google Traduction (n=37), ou sans accès à des aides externes (n=12). Tout en nous invitant à distinguer pratiques effectives et pratiques réelles, les corpus constitués se révèlent complémentaires. En dépit des limites inhérentes à la nature clandestine des pratiques étudiées, ils donnent à voir les difficultés rencontrées par les apprenants de L2 lorsqu'ils consultent les différents outils d'aide à la traduction mais aussi les potentialités de ces différents outils dans une perspective d'apprentissage.

Bourdais, A., & Guichon, N. (2020). Représentations et usages du traducteur en ligne par les lycéens. *Alsic. Apprentissage Des Langues et Systèmes d'Information et de Communication*, *23*.

De Certeau, M. (1980). *L'invention du quotidien. 1. Arts de faire*. Paris : Union générale d'édition.

Hamel, M.-J., Séror, J., & Dion, C. (2015). *Writers in Action : Modelling and Scaffolding Second-Language Learners' Writing Process – Higher Education Quality Council of Ontario*. Toronto : Higher Education Quality Council of Ontario. Retrieved from Higher Education Quality Council of Ontario website: https://heqco.ca/pub/writers-in-action-modelling-and-scaffolding-second-language-learners-writing-process/

Hayes, J., & Flower, L. (1980). Identifying the organization of writing processes. In E. Steinberg & L. Gregg (Eds.), *Cognitive processes in writing : An interdisciplinary approach* (pp. 3–30). Hillsdale, NJ : Lawrence Erlbaum Associates.

**Sarra EL AYARI, Structures Formelles du Langage - UPL, CNRS et Université Paris 8, France**
**Marzena WATOREK, Structures Formelles du Langage - UPL, CNRS et Université Paris 8, France**

## Une plateforme d'annotation pour les corpus écrits d'apprenants

*Mots-clés — exploration de corpus, annotation, interlangue, plateforme, apprenants L2*

Sarramanka est une plateforme d'exploration et d'annotation de corpus, développée pour faciliter l'analyse des corpus écrits et oraux. Il s'agit d'un outil en ligne, qui ne nécessite aucune installation locale. Cet outil ne suppose pas l'adoption d'un format d'annotation pré-établi ; il permet de convertir un schéma d'annotation sous la forme d'un formulaire afin d'annoter les phénomènes finement, ce qui permet plus de rapidité et de fiabilité dans le processus. Le format des annotations ainsi ajoutées peut être adapté en fonction des besoins des chercheurs et chercheuses (CHAT, XML, etc.). Il est également possible de faire des annotations à la volée, et d'ajouter des commentaires afin de créer des sous-corpus d'énoncés. Le corpus peut ensuite être exporté dans différents formats interopérables (EXCEL, XML, CHAT). Une de ses spécificités est de pouvoir naviguer dans l'intégralité du corpus et d'afficher les phénomènes étiquetés avec des jeux de couleurs, qui permettent de visualiser la répartition des phénomes et de mesurer leur proportion très facilement. Il permet ainsi de combiner à la fois des aspects qualitatifs et quantitatifs, sans perdre l'intégralité du corpus de vue.

Cet outil a été testé sur le corpus du projet ESF (Perdue 1993) contenant des productions orales d'apprenants issus de l'immigration, faiblement scolarisés dans leurs langues maternelles et qui acquièrent la L2 en milieu naturel (El Ayari & Watorek 2021). Ce corpus a donné lieu à de nombreuses publications (Perdue 1993, Klein & Perdue 1997) qui ont analysé ces données dans l'approche des lectes d'apprenants issus des approches fonctionnalistes. Les résultats montrent que les productions des apprenants sont une manifestation d'un nouveau système linguistique émergeant d'une interaction entre l'input en L2 et les besoins de communication auxquels l'apprenant doit faire face. Ce système, *lecte des apprenants*, se caractérise par une systématicité interne et peut être décrit par des règles (schémas phrastiques, principes sémantiques et discursives) sous-jacentes à son fonctionnement. Ainsi, les lectes des apprenants ne sont pas une imitation imparfaite de la LC, mais des

systèmes à part entière dépourvus d'erreurs, et à décrire comme une langue inconnue (Klein & Perdue 1997).

Nous présentons les résultats d'une étude exploratoire des productions en français L2 provenant du corpus ESF à l'aide de Sarramanka, qui permet l'exploration, la visualisation et l'annotation de corpus que nous avons développés spécifiquement pour ces données. Aucun outil de ce type n'est disponible à notre connaissance.

De Cock, S. & Tyne, H. (2014). Corpus d'apprenants et acquisition des langues, *Recherches en didactique des langues et des cultures*, 11-1.

El Ayari, S. & Watorek, M. (2021). Exploration outillée pour un corpus de productions orales d'apprenants débutants en L2. Colloque Influence translinguistique : où en est-on aujourd'hui ? Toulouse, France (GIS RéAL2), juillet 2021.

Granger, S. (2004). Computer learner corpus research: current status and future prospects. In Connor, U. & Upton, T. (dir.). Applied corpus linguistics: a multidimensional perspective. Amsterdam/Atlanta:Rodopi. pp. 123-145.

Klein, W. and Perdue, C. (1997). The basic variety (or: couldn't natural languages be much simpler?), Second Language Research 13(4), 301-347.

Perdue, C. (ed.) (1993). Adult Language Acquisition: Crosslinguistic Perspectives. Volume 1, Field Methods. Cambridge: Cambridge University Press.

**Lucia GOMEZ, LIDILEM - Université Grenoble-Alpes, France**

## Triangulation des résultats d'une étude sur corpus d'apprenants (L1/L2) et d'un test expérimental. Le cas de deux verbes polysémiques en espagnol (*poner/volver*)

Les linguistes et les spécialistes de l'enseignement des langues sont unanimes : malgré l'intérêt des différentes études publiées sur les verbes de changement en espagnol, il n'existe actuellement aucune proposition convaincante pour expliquer leur fonctionnement dans la classe d'espagnol L2 (Ibarretxe-Antuñano et Cheikh-Khamis, 2019). Nous avions tenté préalablement de contribuer à la résolution de cette situation en présentant une proposition didactique basée dans les résultats de l'analyse d'un corpus d'apprenants (espagnol L1/L2) (Gómez Vicente 2020). A cette fin, nous avions utilisé le corpus *CALMER* (De Oliveira & Gómez Vicente 2020). Cette approche faisait écho à la problématique soulignée par Granger (2009 14), qui déplorait le fait que les recherches basées sur les corpus d'apprenants soient rarement appliquées à des applications pédagogiques concrètes. Plus concrètement, nous

avions analysé l'utilisation l'usage des verbes 'de changement' *poner* et *volver* en espagnol afin de vérifier si les apprenants L2 d'espagnol ont utilisé ces verbes de la même manière (fréquence, productivité, combinatoire, distribution) que les locuteurs natifs. Pour cette journée d'études, nous allons comparer ces résultats aux résultats obtenus sur un test expérimental, basé dans l'élicitation de phrases liées à ces deux verbes (analyse en cours). La triangulation des données de corpus avec les données du test expérimental permettra de parvenir à une meilleure compréhension du phénomène en question, comme cela a été démontré dans plusieurs études de corpus L2 (Gilquin & Gries 2009 ; Mendikoetxea et Lozano 2018). Ceci nous permettra de donner une nouvelle perspective aux résultats issus de l'étude réalisée sur corpus et éventuellement aussi à la proposition didactique.

Bernardon De Oliveira, Katia & Gómez Vicente. Lucía, (2020). Corpus CALMER [Corpus]. Dans: *ORTOLANG (Open Resources and TOols for LANGuage).* Disponible sur : https://www.ortolang.fr/market/corpora/corpus-calmer/v1

Gilquin, Gaëtanelle &. Gries, Stefan. (2009). Corpora and experimental methods: A state-of-the-art review, *Corpus Linguistics and Lingustic Theory* 5(1).

Gómez Vicente, Lucía. (2020). Aportes de la polisemia para la descripción y la enseñanza de los verbos de cambio. Análisis del uso de *poner* y *volver* en l1 y l2*. RLA. Revista De Lingüística Teórica Y Aplicada*, 58(2)

Granger, Sylviane. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. Dans: Aijmer, Karin (Ed.), *Corpora and Language Teaching*, Benjamins: Amsterdam and Philadelphia.

Ibarretxe-Antuñano, Iraide, Cheikh-Khamis, Fátima. (2019). How to become a woman without turning into a Barbie: Change-of-state verb constructions and their role in Spanish as a Foreign Language. *International Review of Applied Linguistics* (*IRAL*) 57(1).

Lozano, Cristóbal & Mendikoetxea, Amaya. (2013). Corpus and experimental data: Subjects in second language research. En Granger, Sylviane, Gilquin Gaëtannelle & Meunier, Fanny (Eds.), *Twenty Years of Learner Corpus Research: Looking back, Moving ahead. Corpora and Language in Use – Proceedings* 1, Louvain-la-Neuve: Presses universitaires de Louvain.

**Qianyun LI, LIDILEM - Université Grenoble Alpes, France**

## Le traitement des données contrastives en langues éloignées

*Mots-clés — Analyse contrastive, glose, corpus bilingue*

Dans le cadre de l'analyse contrastive, les productions de locuteurs natifs dans leur langue première (L1) constituent un corpus de référence pour expliquer les difficultés particulières des textes d'apprenants en langue étrangère (LE). De fait, les erreurs commises résultent souvent de différences entre la L1 et la LE. Les apprenants ont tendance à transférer des

éléments de leur L1 lors de la production écrite en LE, ce qui rend la compréhension de leurs textes plus difficile pour les locuteurs natifs (Wang & Wen, 2002). Toutefois, la nécessité de comparer des corpus bilingues ou multilingues soulève une interrogation méthodologique : comment traiter les données dans les langues typologiquement éloignées ?

Cette communication se propose de présenter et mettre en perspective une méthodologie de transcription appliquée à l'acquisition du français. À la différence du français représenté en l'alphabet latin, le chinois appartenant aux familles sino-tibétaines se compose de sinogrammes. Avant de passer à l'analyse des données, nous envisagerons de transcrire les textes en sinogrammes produits par des sinophones natifs.

Nous proposons d'utiliser la règle de *Grammatical category labels* (Department of Linguistics, 2015) pour transcrire des textes en chinois. En effet, le développement de *Leipzig Glossing Rules* (*ibid*.) fournit une liste des conventions en matière de transcription et de description des données bilingues. Cependant, la façon de gloser est associée à l'intention de l'auteur et aux besoins du public fixé. Dans l'objectif d'analyser l'organisation textuelle, nous tenterons donc de gloser des phrases en chinois aux niveaux syntaxique et sémantique. Ceci implique de diviser chaque phrase en mots simples ou en mots composés porteurs de signification. La traduction littérale des mots et l'abréviation de leur catégorie grammaticale sont présentées comme suit, en deux lignes successives :

| 文凭 | 就 | 是 | 学历, | 是 | 求 | 职 | 的 | 敲门砖。 |
|---|---|---|---|---|---|---|---|---|
| diplôme | exactement | être | niveau d'études | être | chercher | travail | DE | moyen d'entrée |
| NS | Adv. | V | NS | V | V | NS | PS | NS |
| Le diplôme est exactement le niveau d'études, (il) est un moyen d'entrée de chercher un travail. | | | | | | | | |

Cette glose des données facilite la lisibilité et l'analyse des textes en chinois vis-à-vis de lecteurs non sinophones. À travers certains exemples glosés, le lecteur peut saisir le contenu du texte et manipuler les arguments linguistiques de scripteurs (Lehmann, 1982). La glose des catégories grammaticales sous forme d'abréviation nous semble pertinente car elle permet au lecteur de comprendre la structure grammaticale et le sens des données en L1 sans translation contrôlée, et sans ajout d'informations complémentaires.

Department of Linguistics. (2015). *Conventions for interlinear morpheme-by-morpheme glosses*. https://www.eva.mpg.de/lingua/resources/glossing-rules.php

Lehmann, C. (1982). Directions for interlinear morphemic translations. *Folia Linguistica*, *16*, 199-224.

Wang, W., & Wen, Q. (2002). L1 Use in the L2 Composing Process : An Exploratory Study of 16 Chinese EFL Writers. *Journal of Second LanguageWriting*, *11* (3), 225-246.

**Xinyue Cécilia YU, CRLAO – Institut National des Langues et Civilisations Orientales (Inalco), France**
**Pierre MAGISTRY, ERTIM – Institut National des Langues et Civilisations Orientales (Inalco), France**

## Constitution d'un corpus écrits d'apprenants du chinois langue étrangère - Premières réflexions et démarches entreprises

*Mots-clés — corpus écrit d'apprenants, L1 français, L2 chinois*

L'exploitation des corpus d'apprenants dans les recherches en acquisition des langues étrangères (RALE) a connu un développement important en France, principalement pour des études en français langue étrangère et anglais langue étrangère. En revanche, dans le domaine du chinois langue étrangère, la constitution des corpus, à notre connaissance, reste *ad hoc* et sur objectif spécifique.

Afin d'avoir une perspective panoramique et de suivre le processus d'acquisition (Meunier, 2019 : 34-44) des apprenants francophones du chinois langue étrangère, nous sommes en train de construire un corpus écrit d'apprenants en contexte universitaire. Celui-ci couvre 4 niveaux différents du LLCER chinois. Le présent travail résume nos premières réflexions et étapes dans notre projet.
Notre objectif principal est de collecter les rédactions écrites des apprenants tout au long de l'année universitaire afin de constituer un corpus qui nous permettra d'analyser le processus d'acquisition du niveau débutant jusqu'au niveau avancé sur les plans lexical, grammatical et textuel.

La première démarche entreprise dans la construction du corpus consiste en l'élaboration d'un formulaire de consentement au recueil de données, que les enseignants de l'équipe ont fait signer aux étudiants qui étaient d'accord pour participer à notre étude. Les apprenants volontaires sont au nombre de 50 à 120 en fonction du niveau d'études. Ce qui nous permet de recueillir les productions écrites d'au moins 250 apprenants en une année (compte tenu des désistements éventuels).

Depuis la rentrée 2021, des espaces de dépôt de document ont été créés sur la plateforme Moodle pour que les étudiants, avec l'aide des enseignants, mettent en ligne au fur et à mesure leurs rédactions. Les productions comprennent 3 à 5 devoirs à la maison et 1 ou 2 devoirs sur table par semestre. Ainsi, les rédactions collectées pourront servir d'une part pour des études longitudinales d'une année pour chaque niveau, et d'autre part pour des études pseudo-longitudinales en comparant les productions des différents niveaux (Guilquin, 2015 : 9-34).

À partir des premiers manuscrits collectés, nous avons commencé à travailler sur la grille d'annotation d'erreur (Lüdeling & Hirschmann, 2015 : 135-157). Nous proposons d'axer notre présentation sur la méthodologie du recueil des données et l'analyse préliminaire des premières données collectées.

Gilquin, G., & Granger, S. (2015). From design to collection of learner corpora. *The Cambridge handbook of learner corpus research*, *3*(1), 9-34.

Lüdeling, A., & Hirschmann, H. (2015). Error annotation systems. *The Cambridge handbook of learner corpus research*, 135-157.

Meunier, F. (2019). Tracking developmental patterns in learner corpora: Focus on longitudinal studies. *Selected papers on theoretical and applied linguistics*, *23*, 34-44.