La collection Pangloss, une archive ouverte de langues rares

Alexis Michaud et Séverine Guillaume

La collection Pangloss (pangloss.cnrs.fr) est une archive ouverte de langues rares, membre du réseau international DELAMAN (Digital Endangered Languages and Musics Archive Network). La gestion en est assurée par le LACITO (« Langues et Civilisations à Tradition Orale »), unité mixte de recherche (CNRS-Université Sorbonne Nouvelle-Inalco) qui a pour mission d'explorer et sauvegarder la diversité des langues et des traditions orales, dans un monde où cette diversité décroît rapidement.

Une archive ouverte

Se constituer un butin de paroles gelées, « dragées perlées de diverses couleurs », est devenu bien facile ; mais comme l'imaginait Rabelais (*Quart Livre*, LV), ces capsules que l'on peut réécouter à l'envi, on ne peut pour autant les entendre, « car c'est langage barbare ». Les outils de l'ère numérique permettent de constituer des ressources multimédia qui combinent les avantages de l'écrit (sous forme d'annotations structurées logiquement) et ceux de la parole : conserver l'étoffe d'une voix, dans toute sa fraîcheur, et restituer son contexte social, historique et linguistique. Les documents sont consultables en ligne, au moyen d'un simple navigateur web, sans qu'il soit besoin de télécharger de logiciel dédié. L'interface web est maintenue et améliorée au fil du temps, au rythme de l'évolution des outils et du design web. Les choix techniques pour le format des documents, en revanche, sont pour l'essentiel inchangés depuis les débuts de l'archive il y a plus de vingt ans. Ils sont exposés dans diverses publications auxquelles on se permet de renvoyer ici (Jacobson, Michailovsky & Lowe 2001 ; Michailovsky et al. 2014 ; Vasile et al. 2020). En résumé : des annotations au format XML (transcription, traduction, gloses, commentaires) sont associées, dans la mesure du possible, aux fichiers audio et vidéo. En termes de langues, *Pangloss* n'est, comme son nom l'indique, fermé à aucune langue : « langues rares » est compris dans un sens large, qui coïncide en pratique avec les objets de recherche des linguistes « de terrain ». Il n'est pas posé de limite, inférieure ni supérieure, à la taille d'un corpus accepté pour dépôt : les données de chercheuses et chercheurs sont accueillis dans l'état de traitement auquel elles ont été portées par les déposant·es, suivant le principe selon lequel quelques données, même très incomplètes, valent déjà infiniment mieux que l'absence de toute donnée. Parmi les 250 corpus que comporte actuellement la collection (en 200 langues environ), certains ne comptent que quelques minutes d'enregistrements non transcrits, tandis que le plus étendu comporte plusieurs dizaines d'heures d'enregistrements transcrits. Le périmètre de l'archive est national : les déposants doivent être des acteurs de l'enseignement supérieur et de la recherche français (ou des partenaires de leurs travaux).

Un important potentiel en recherche

On aimerait souligner ici le potentiel que présente la collection Pangloss pour les linguistes. Limiter la perte des données récoltées lors de missions de terrain constitue certes un enjeu primordial, s'agissant de langues rares et en danger. Mais l'archivage pérenne des données de la recherche n'est pas la fin dernière du projet. Il ne s'agit en effet pas de thésauriser des fichiers, mais d'ouvrir les corpus à une exploitation renouvelée en recherche, y compris avec les outils de la linguistique de corpus.

Un élément d'un écosystème de Science ouverte

La collection Pangloss est pionnière du mouvement qu'on appelle aujourd'hui de Science ouverte, et qui n'est autre qu'une façon de travailler guidée par l'application conséquente de principes scientifiques simples et de bon sens. La collection a par là vocation à être un lieu de réflexion sur les usages des données, et sur le processus d'enrichissement continu des ressources après qu'elles aient rejoint le *jardin de données* que constitue l'archive ouverte, mais aussi sur les dispositifs de collecte (dans les localités où la parole *coule de source*), en dialogue avec les communautés concernées.

Une réflexion au sujet du Traitement Automatique des Langues

Le dernier point qu'on soulignera ici est que la collection Pangloss vise à faciliter une utilisation des jeux de données en Traitement Automatique des Langues, dans le cadre de partenariats équilibrés. L'équipe de la collection Pangloss participe au développement d'outils numériques, pour faire fonctionner et enrichir la collection elle-même, mais également pour faire avancer la recherche au moyen des technologies de la parole : outils de transcription automatique de la parole, notamment (on mentionnera en particulier : Wisniewski et al. 2020 ; Adams et al. 2021 ; Guillaume et al. 2022 ; Guillaume, Wisniewski & Michaud 2023 ; Fily et al. 2024).

Ces quelques mots de présentation de la collection Pangloss ne prétendent à l'évidence en aucune façon à aborder l'ensemble des questions liées à la constitution, l'archivage et l'exploitation de ressources en langues rares. L'état actuel de la collection, les liens avec l'enseignement et la médiation scientifique, les questions éthiques et juridiques, aussi bien que le volet technique, mériteraient de longs développements. L'équipe de la collection Pangloss recevrait commentaires, suggestions et propositions de collaboration avec un vif intérêt.

Références citées :

Adams, Oliver, Benjamin Galliot, Guillaume Wisniewski, Nicholas Lambourne, Ben Foley, Rahasya Sanders-Dwyer, Janet Wiles, et al. 2021. User-friendly automatic transcription of low-resource languages: plugging ESPnet into Elpis. In *Proceedings of ComputEL-4: Fourth Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Hawai'i. https://halshs.archives-ouvertes.fr/halshs-03030529.

Fily, Maxime, Guillaume Wisniewski, Severine Guillaume, Gilles Adda & Alexis Michaud. 2024. Establishing degrees of closeness between audio recordings along different dimensions using large-scale cross-lingual models. arXiv. http://arxiv.org/abs/2402.05581. (20 February, 2024).

Guillaume, Séverine, Guillaume Wisniewski, Benjamin Galliot, Minh-Châu Nguyễn, Maxime Fily, Guillaume Jacques & Alexis Michaud. 2022. Plugging a neural phoneme recognizer into a simple language model: a workflow for low-resource settings. In *Proceedings of Interspeech 2022*. Incheon, Korea. https://halshs.archives-ouvertes.fr/halshs-03625581.

Guillaume, Séverine, Guillaume Wisniewski & Alexis Michaud. 2023. From 'snippet-lects' to doculects and dialects: Leveraging neural representations of speech for placing audio signals in a language landscape. In *Proceedings of SIGUL 2023: 2nd Annual Meeting of the Special Interest Group on Under-resourced Languages*. Dublin. https://hal.science/hal-04108652.

Jacobson, Michel, Boyd Michailovsky & John B. Lowe. 2001. Linguistic documents synchronizing sound and text. *Speech Communication* 33 [special issue: "Speech Annotation and Corpus Tools"]. 79–96.

Michailovsky, Boyd, Martine Mazaudon, Alexis Michaud, Séverine Guillaume, Alexandre François & Evangelia Adamou. 2014. Documenting and researching endangered languages: the Pangloss Collection. *Language Documentation and Conservation* 8. 119–135.

Vasile, Aurelia, Séverine Guillaume, Mourad Aouini & Alexis Michaud. 2020. Le Digital Object Identifier, une impérieuse nécessité ? L'exemple de l'attribution de DOI à la Collection Pangloss, archive ouverte de langues en danger. *I2D - Information, données & documents* 2. 156–175.

Wisniewski, Guillaume, Alexis Michaud, Benjamin Galliot, Laurent Besacier, Séverine Guillaume, Katya Aplonova & Guillaume Jacques. 2020. Ouvrir aux linguistes « de terrain » un accès à la transcription automatique. In *Actes des Journées scientifiques du Groupement de Recherche "Linguistique informatique, formelle et de terrain" (LIFT)*. Paris. https://hal.archives-ouvertes.fr/hal-03047148.