Deux corpus audio transcrits de langues rares (japhug et na) normalisés en vue d'expériences en traitement du signal

Benjamin Galliot, Guillaume Wisniewski, Séverine Guillaume, Laurent Besacier, Guillaume Jacques, Alexis Michaud, Solange Rossato, Minh-Châu Nguyễn, Maxime Fily

Deux corpus audio transcrits de langues rares (japhug et na) normalisés en vue d'expériences en traitement du signal

Benjamin Galliot[2] Guillaume Wisniewski[3] Séverine Guillaume[2]
Laurent Besacier[4] Guillaume Jacques[5] Alexis Michaud[2] Solange Rossato[1]
Minh-Châu Nguyên[1, 2] Maxime Fily[2]

(1) Laboratoire d'Informatique de Grenoble (LIG), Unité Mixte de Recherche 5217 CNRS - Université Grenoble Alpes - Grenoble INP - Institut national de recherche en informatique et en automatique (INRIA)
(2) Langues et Civilisations à Tradition Orale (LACITO), Unité Mixte de Recherche 7107 CNRS - Sorbonne Nouvelle - Institut National des Langues et Civilisations Orientales (INALCO)
(3) Laboratoire de Linguistique Formelle (LLF), Unité Mixte de Recherche 7110 CNRS - Université de Paris
(4) Naver Labs Europe, Grenoble
(5) Centre de Recherches Linguistiques sur l'Asie Orientale (CRLAO), Unité Mixte de Recherche 8563 CNRS - École des Hautes Études en Sciences Sociales - Institut National des Langues et Civilisations Orientales
b.g01lyon@gmail.com, Guillaume.Wisniewski@univ-paris-diderot.fr, severine.guillaume@cnrs.fr, laurent.besacier@univ-grenoble-alpes.fr, rgyalrongskad@gmail.com, alexis.michaud@cnrs.fr, Solange.Rossato@univ-grenoble-alpes.fr, minhchau.ntm@gmail.com, maxime.fily@gmail.com

RÉSUMÉ

Deux corpus audio transcrits de langues « rares » (langues minoritaires de Chine : japhug et na) sont proposés comme corpus de référence pour des expériences en traitement automatique des langues. Les données, collectées et transcrites au fil d'enquêtes de terrain en immersion, s'élèvent à un total de 1907 minutes d'audio transcrit en japhug et de 209 minutes en na. Nous décrivons les traitements effectués pour les mettre à disposition sous une forme aisément accessible et utilisable, et présentons un outil qui permet d'assembler divers jeux de données de la collection Pangloss (archive ouverte de langues rares) en assurant la reproductibilité des expériences menées sur ces données.

ABSTRACT

Two Very-Low-Resource Language Speech Corpora for Experiments in NLP : Japhug and Na

Two audio corpora of minority languages of China (Japhug and Na), with transcriptions, are proposed as reference data sets for experiments in Natural Language Processing. The data, collected and transcribed in the course of immersion fieldwork, amount to a total of 1,907 minutes in Japhug and 209 minutes in Na. By making them available in an easily accessible and usable form, we hope to facilitate the development and deployment of state-of-the-art NLP tools for the full range of human languages. We present a tool for assembling datasets from the Pangloss Collection (an open archive) in a way that ensures full reproducibility of experiments conducted on these data.

MOTS-CLÉS : Corpus de référence, documentation computationnelle des langues, langues rares.

KEYWORDS: Benchmark datasets, Computational Language Documentation, low-resource languages, endangered languages.

1 Introduction

Le déploiement d'outils de traitement automatique de la parole comporte des enjeux évidents pour la documentation des langues, à une époque où le déclin de la diversité linguistique s'accélère (parallèlement au déclin de la biodiversité). Inversement, les langues rares présentent à la recherche en informatique tout un éventail de défis dont l'intérêt est de plus en plus clairement perçu.

Dans ce contexte, la mise à disposition de corpus de langues rares aisément accessibles, clairement versionnés et faciles d'utilisation paraît une nécessité tout à fait centrale. Dans le droit fil de la publication du corpus mbochi (bantou), décrite par Godard et al. (2018), nous avons déposé dans Zenodo deux corpus audio (avec transcriptions) de langues rares : le japhug et le na, langues minoritaires de Chine, de la famille sino-tibétaine.

Ces corpus ont été utilisés dans des travaux innovants en reconnaissance automatique de la parole (Adams et al., 2018, 2021; Macaire, 2021) et dans des réflexions interdisciplinaires associant talistes et linguistes (Michaud et al., 2018, 2019, 2020). Les corpus sont disponibles en ligne dans la collection Pangloss [1] (Michaud et al., 2016), une archive ouverte de langues rares hébergée par la plateforme Cocoon [2], mais les transcriptions de ces corpus sont enrichies et revues au fil des années, et de nouveaux documents s'y ajoutent, de sorte que renvoyer simplement au *corpus de langue L dans Pangloss* ne constitue pas une référence suffisamment précise pour parvenir à une reproductibilité d'expériences de TAL (ou d'autre type) menées sur ces données. Il s'agit, sinon de « données chaudes », du moins de « données tièdes », qui évoluent lentement au fil du temps.

Nous avons donc effectué un dépôt dans Zenodo d'un état donné de ces deux corpus, ainsi rendu accessible en quelques clics, sous une forme stabilisée (§2). Pour les collègues qui souhaiteraient aller au-delà d'une utilisation en l'état de ces deux jeux de données proposés au statut de *corpus de référence*, un outil est proposé (§3) pour panacher à son aise parmi l'ensemble des collections tout en conservant une garantie de reproductibilité, grâce au système de versionnage dont bénéficient les documents de la plateforme Cocoon.

2 Les dépôts : liens d'accès et choix techniques

Lieu de dépôt Les deux corpus ont été téléversés sur Zenodo, dont ils constituent respectivement les dépôts 5336698 (na) et 5521112 (japhug). Un corpus entier est identifié par un DOI (*digital object identifier*) : `10.5281/zenodo.5336698` pour le na, et `10.5281/zenodo.5521112` pour le japhug.

Le même type d'identifiant a été déployé pour la collection Pangloss, l'archive ouverte où sont déposés les corpus, mais avec une granularité tout à fait différente : un DOI pour chaque document (Vasile et al., 2020), ce qui est bien adapté pour les linguistes qui souhaitent faire référence aux données avec une granularité fine (un texte et, à l'intérieur d'un texte, un énoncé précis) mais ne donne pas prise sur un corpus entier.

Fichiers audio Les fichiers audio ont été dégradés en 16 bit, 16 kHz, mono. Là aussi, la logique qui préside à la constitution de ces corpus versionnés pour expériences de TAL s'éloigne de celle

1. `https://pangloss.cnrs.fr/`
2. `https://cocoon.huma-num.fr/`

de l'archivage pérenne dans la collection Pangloss. La taille des deux jeux de données déposés dans Zenodo est compatible (au jour d'aujourd'hui) avec des expériences menées sur un ordinateur portable : 1,8 Go pour le na, 9,2 Go pour le japhug.

Annotations Les annotations sont dans le format d'origine : du XML organisé selon une hiérarchie simple (un texte est composé de phrases, composées de mots, composés de morphèmes). Un prétraitement élémentaire a été effectué, afin de ne pas imposer aux utilisateurs de devoir prendre connaissance d'un certain nombre de conventions choisies par les déposants. En particulier, lors de la transcription de textes, il arrive que des retouches soient apportées, qui éloignent la transcription de ce qui a été dit sur l'enregistrement ; les passages ajoutés sont signalés par des crochets `[]`, et les passages que les consultants linguistiques souhaitent voir retranchés de la transcription « lissée » sont placés entre chevrons <>. Lors du prétraitement, les premiers ont été effacés, et les seconds allégés de leurs chevrons, afin qu'audio et transcription coïncident.

3 Un outil pour constituer de nouveaux jeux de données

Au-delà des deux jeux de données déposés dans Zenodo, il est bien sûr possible de constituer toutes sortes de nouveaux jeux de données à partir de la collection Pangloss. L'outil élaboré à l'occasion de la préparation des deux corpus déposés dans Zenodo, sobrement intitulé OutilsPangloss [3], consiste en une boîte à outils divers (en langage Julia) servant notamment à créer des (sous-)corpus de langues rares de Pangloss.

L'utilisateur·rice remplit un fichier YAML (dont différents exemples sont fournis), en indiquant notamment le nom de la langue. Elle peut également fournir une liste d'expressions rationnelles de modifications si des traitements sur les annotations sont à faire (telles que suppressions ou réarrangements de blocs de textes). Il est possible de filtrer par locuteur pour les sous-corpus. Des traitements sur l'audio peuvent également être paramétrés, pour choisir le taux d'échantillonnage et la profondeur, et séparer les différentes pistes des fichiers multicanaux en fichiers mono (démultiplexage).

Après le moissonnage (en Sparql), les vérifications des données par les métadonnées (hachage, versions, etc.) et les téléchargements, un fichier récapitulatif général (`données.yml`) se trouvera dans le dossier cible, aux côtés de dossiers `données` et `métadonnées`. Les informations contenues dans ce fichier récapitulatif suffisent à reproduire exactement, à tout moment, une expérience menée avec le jeu de données qu'il décrit, de sorte qu'il n'est pas techniquement nécessaire de travailler avec le jeu de données lui-même dans Zenodo.

Paradoxalement, si l'objet premier de la présente communication est d'attirer l'attention vers des jeux de données mis à disposition dans Zenodo, notre espoir serait que les pratiques s'orientent à l'avenir vers une description des jeux de données via des métadonnées renvoyant vers un unique hébergement des données dans une archive pérenne. En effet, décrire de cette façon le jeu de données qu'on a utilisé ne prend que quelques kilooctets (ko), tandis qu'un dépôt *en dur* de chaque bouquet de données multiplierait des dépôts (dans Zenodo ou ailleurs) dont chacun se compte en gigaoctets (Go).

On se contentera, en guise de mot de la fin, de relever que c'est aux différents acteurs de nos domaines de recherche qu'il appartient de s'emparer de ces outils, et de contribuer à façonner, par leurs choix, les directions que prendront les pratiques à l'interface du TAL et de la documentation des langues rares, dans le contexte d'une transition en cours vers la science ouverte.

3. `https://gitlab.com/lacito/outilspangloss`

Remerciements

Un grand merci aux collègues et amis consultants de langue japhug (en particulier Tshendzin) et na (en particulier M^{me} Latami Dashilame et son fils Latami Dashi). Le présent travail est une contribution au projet « La documentation computationnelle des langues à l'horizon 2025 » (ANR-19-CE38-0015-04) ainsi qu'au Labex « Fondements empiriques de la linguistique » (ANR-10-LABX-0083). Nous remercions l'Institut des langues rares (ILARA) de l'École pratique des hautes études, l'Université du Queensland et l'*Australian Research Council Centre of Excellence for the Dynamics of Language* pour le soutien financier apporté au développement d'outils logiciels pour la documentation linguistique.

Références

Adams, O., Cohn, T., Neubig, G., Cruz, H., Bird, S., and Michaud, A. (2018). Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of LREC 2018 (Language Resources and Evaluation Conference)*, pages 3356–3365, Miyazaki.

Adams, O., Galliot, B., Wisniewski, G., Lambourne, N., Foley, B., Sanders-Dwyer, R., Wiles, J., Michaud, A., Guillaume, S., Besacier, L., Cox, C., Aplonova, K., Jacques, G., and Hill, N. (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.

Godard, P., Adda, G., Adda-Decker, M., Benjumea, J., Besacier, L., Cooper-Leavitt, J., Kouarata, G. N., Lamel, L., Maynard, H., and Mueller, M. (2018). A very low resource language speech corpus for computational language documentation experiments. In *Proceedings of LREC 2018 (Language Resources and Evaluation Conference)*, pages 3366–3370, Miyazaki.

Macaire, C. (2021). Recognizing lexical units in low-resource language contexts with supervised and unsupervised neural networks. Master's thesis, Université de Lorraine.

Michaud, A., Adams, O., Cohn, T., Neubig, G., and Guillaume, S. (2018). Integrating automatic transcription into the language documentation workflow : experiments with Na data and the Persephone toolkit. *Language Documentation and Conservation*, 12 :393–429. Dans HAL : https ://halshs.archives-ouvertes.fr/halshs-01841979/.

Michaud, A., Adams, O., Cox, C., and Guillaume, S. (2019). Phonetic lessons from automatic phonemic transcription : preliminary reflections on Na (Sino-Tibetan) and Tsuut'ina (Dene) data. In *Proceedings of ICPhS XIX (19th International Congress of Phonetic Sciences)*, Melbourne.

Michaud, A., Adams, O., Cox, C., Guillaume, S., Wisniewski, G., and Galliot, B. (2020). La transcription du linguiste au miroir de l'intelligence artificielle : réflexions à partir de la transcription phonémique automatique. *Bulletin de la Société de Linguistique de Paris*, 116(1).

Michaud, A., Guillaume, S., Jacques, G., Mac, D.-K., Jacobson, M., Pham, T.-H., and Deo, M. (2016). Contribuer au progrès solidaire des recherches et de la documentation : la Collection Pangloss et la Collection AuCo. In *Journées d'Etude de la Parole 2016*, volume 1, pages 155–163.

Vasile, A., Guillaume, S., Aouini, M., and Michaud, A. (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.