

Faciliter l'accès des praticiens du Traitement Automatique des Langues à des jeux de données de langues rares : un deuxième point d'étape

Faciliter l'accès des praticiens du Traitement Automatique des Langues à des jeux de données de langues rares : un deuxième point d'étape

Benjamin Galliot¹ Guillaume Wisniewski² Séverine Guillaume¹
Guillaume Jacques³ Alexis Michaud¹

(1) Langues et Civilisations à Tradition Orale (LACTO), CNRS - Sorbonne Nouvelle - INALCO
(2) Laboratoire de Linguistique Formelle (LLF), CNRS - Université de Paris
(3) Centre de Recherches Linguistiques sur l'Asie Orientale (CLIAO), CNRS - Ecole des Hautes Études en Sciences Sociales - Institut National des Langues et Civilisations Orientales
b.galliot@gmail.com, Guillaume.Wisniewski@univ-paris-diderot.fr, severine.guillaume@cnrs.fr, rgyalrongskad@gmail.com, alexis.michaud@cnrs.fr

RÉSUMÉ

Nous présentons un outil logiciel qui permet d'assembler divers jeux de données de la collection Pangloss (archive ouverte multimédia de langues rares) en assurant la reproductibilité des expériences menées sur ces données. À titre d'exemple, deux corpus audio transcrits de langues minoritaires de Chine (japhug et na) sont proposés, sous une licence Creative Commons, comme corpus de référence pour des expériences en traitement automatique des langues, et comme exemples d'une chaîne de traitement généralisable à d'autres corpus d'archives ouvertes. L'enjeu global d'une mise à disposition de données de langues rares sous une forme aisément accessible et utilisable est de faciliter le développement et le déploiement d'outils de pointe en traitement automatique des langues naturelles pour tout l'éventail des langues humaines. Cet exposé, qui fait suite à une précédente communication sur le même thème, fait état de nouveautés dont un retour d'expérience concernant un dépôt auprès de Hugging Face.

ABSTRACT

Facilitating NLP specialists' access to language archive materials : an update

We present a software tool to assemble a great range of diverse datasets from the Pangloss collection (a multimedia open archive of under-documented languages). The tool ensures the reproducibility of experiments conducted on these data. As an example, two transcribed audio corpora of Chinese minority languages (Japhug and Na) are proposed, under a Creative Commons license, as reference corpora for experiments in Natural Language Processing, and as examples of a pipeline that can be generalized to other corpora from open archives. An overarching goal of making language archive data available in an easily accessible and usable form is to facilitate the development and deployment of state-of-the-art natural language processing tools for the full range of human languages. This presentation, which follows a previous paper on the same topic, reports on new developments including feedback on a deposit at Hugging Face Datasets.

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

KEYWORDS : Benchmark datasets, Computational Language Documentation, endangered languages.

Difficulté identifiée : accès des TAListes aux corpus de langues rares. Archive ouverte Cocoon (Collection de Corpus Oraux Numériques) : pour consultation en ligne plutôt que téléchargement en masse.

The screenshot shows the Pangloss website interface. At the top, there's a navigation bar with 'Qui sommes-nous?', 'Corpus', 'Dictionnaires', and 'Contact'. The main content area is titled 'Bienvenue dans la collection Pangloss, archive ouverte de langues en danger et sous-documentées.' Below this, there's a section 'Ressources en vedette' featuring a video thumbnail for 'Le voleur et le roi' with the subtitle 'khroskyabs Chine'. To the right, there's a search bar and a list of corpus items. One item is highlighted: 'Souvenir d'enfance : jeûner pendant la moisson'. Below the search bar, there are options for 'Options d'affichage' and 'Lecture en continu'. The interface is clean and modern, with a focus on accessibility and ease of use.

Besoin de reproductibilité des expériences. Exemple d'utilisation : explorations en Reconnaissance Automatique de la Parole.

Hugging Face Datasets : disponible pour expériences



The screenshot shows the Hugging Face Datasets interface. The search bar contains 'pangloss'. Below the search bar, there are filters for 'Tasks' (Automatic Speech Recognition, Fine-Grained Tasks), 'Languages' (jya, nru), 'Multilinguality' (multilingual, translation), and 'Size Categories'. The 'Dataset card' for 'pangloss' is displayed, showing the dataset name, a subset of 'yong1288', and a table of data points. The table has columns for 'path (string)', 'audio (audio)', 'sentence (string)', 'doctype (string)', and 'translation:fr (string)'. The first row shows a path starting with 'cocoon-d62e5852-a63f-3674-b09c-6b175a21cb81_C1/Tone_BodyParts01Animals_6_Ve...', an audio player, a sentence in a minority language, the doctype 'WORDLIST', and the translation 'intestin de poulain'.

Zenodo : téléchargement en quelques clics. Formats : selon pratiques courantes en traitement du signal (audio dégradé en 16 kHz, 16-bit) et du texte (transcriptions et annotations en XML ; simplifications mineures pour confort d'utilisation

The screenshot shows the Zenodo website interface. The search bar contains 'Japhug for Natural Language Processing'. Below the search bar, there are filters for 'Video/Audio' and 'Open Access'. The 'Record card' for 'Japhug for Natural Language Processing: a single-speaker audio corpus with transcriptions' is displayed, showing the record title, the date 'September 22, 2021', the number of views (30) and downloads (5), and the DOI '10.5281/zenodo.5521112'. The record is indexed in OpenAIRE.

Outil générique pour création de jeux de données intégralement documentés : OutilsPangloss

Ce programme sobrement intitulé **OutilsPangloss** consiste en une boîte à outils divers servant notamment à créer des (sous-)corpus de langues rares de **Pangloss**.

Installation

Ce programme est codé en Julia et téléchargera si besoin est un moteur XSLT 2/3 (Saxon), ainsi Java devrait aussi être installé sur l'ordinateur hôte. Du fait de sa nature, il est nécessaire d'avoir une connexion internet.

Utilisation

Création de corpus

Tout ce dont a besoin un utilisateur est de remplir un fichier YAML (dont différents exemples se trouvent dans le dossier `exemples`) dont les éléments importants sont les suivants :

- le nom de la langue du corpus qui se trouve sur Pangloss (« Japhug », « Yongning Na », par exemple) ;
- la liste des graphèmes (notamment complexes) ;
- la liste des expressions rationnelles de modifications si des traitements sur les annotations sont à faire (suppressions ou réarrangements de blocs de textes...) ;
- les informations sur les corpus et sous-corpus, notamment :
 - le filtre de locuteur pour le sous-corpus ;
 - la langue du fichier récapitulatif associé (français : fr ou anglais : en) ;
 - les traitements à réaliser sur l'audio, notamment :
 - le taux d'échantillonnage (typiquement 16000 Hz) ;

This archive contains a dataset (audio files and transcriptions) of a minority language, Yongning Na (iso 639-3 code: nru). The archive contains a subset of the Na corpus of the Pangloss Collection: it is a single-speaker corpus, consisting of all the audio resources transcribed, for the main speaker of this corpus (Ms. LATAMI Dashilame). The corpus is versioned, so that the experiments carried out on these resources (for linguistic research or for Natural Language Processing) are fully reproducible. All relevant information is contained in YAML files (yml extension; one in French, one in English). The data sub-folder contains the converted and demultiplexed audio files, as well as the annotations associated with each channel of the audio files. The summary files contain, among other things, the list of graphemes used in the language (complex graphemes are particularly important), as well as information on the various resources (audio and annotations), such as their identifiers (DOIs) and links to the original files. From a computational point of view, the list of DOIs of the audios and annotations described in this YAML file is sufficient to generate this corpus at a given time. A corpus like the present one can be viewed as the version, at a given time, of a set of documents in the Pangloss collection: a corpus as it stands at a precise version. Further information is available from <https://github.com/lacito/outlispangloss> (<https://github.com/lacito/outlispangloss>)

Cette archive contient un jeu de données (audios et transcriptions) d'une langue à tradition orale, le na de Yongning (code iso 639-3 : nru). L'archive contient un sous-ensemble du corpus na de la collection Pangloss : c'est un corpus monolocuteur, constitué de l'intégralité des ressources audio transcrites pour la locutrice principale de ce corpus (Mme LATAMI Dashilame). Le corpus est versionné, de sorte que les expériences menées sur ces ressources (pour la linguistique ou pour le Traitement automatique des langues) soient reproductibles de façon exacte (en pensant bien à joindre l'algorithme : paramètres, répartitions des fichiers dans les différents ensembles, etc.). Toutes les informations pertinentes se trouvent dans les fichiers YAML (extension .yml : un en français, un autre en anglais). Le sous-dossier des données contient d'une part les audios convertis et démultiplexés et d'autre part les annotations associées à chaque canal des audios. Les fichiers récapitulatifs contiennent notamment la liste des graphèmes utilisés dans cette langue (les graphèmes complexes sont particulièrement importants), ainsi que des informations sur les différentes ressources (audios et annotations), comme les identifiants (DOI), les liens vers les fichiers originaux, etc. Au plan informatique, la liste des identifiants DOI des audios et annotations décrits dans ce fichier YAML suffit pour générer ce corpus à un instant t. Un corpus comme celui-ci peut être vu comme la version à l'instant t d'un ensemble de documents de la collection Pangloss : un corpus arrêté à une version précise. Pour plus de précisions : <https://github.com/lacito/outlispangloss> (<https://github.com/lacito/outlispangloss>)

The screenshot shows the Zenodo file list for the 'Japhug - Tshendzin - 16k-16.zip' dataset. The list contains 16 files, including audio files (wav) and XML files (xml). The files are organized into a tree structure, with the root folder being 'Japhug - Tshendzin - 16k-16'. The files include audio files for different speakers and transcriptions, as well as XML files for annotations and metadata. The total size of the files is 9.2 GB.