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 Galliot[1] Guillaume Wisniewski[2] Séverine Guillaume[1]
Guillaume Jacques[3] Alexis Michaud[1]
(1) Langues et Civilisations à Tradition Orale (LACITO), CNRS - Sorbonne Nouvelle - INALCO
(2) Laboratoire de Linguistique Formelle (LLF), CNRS - Université de Paris
(3) Centre de Recherches Linguistiques sur l’Asie Orientale (CRLAO), 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, 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

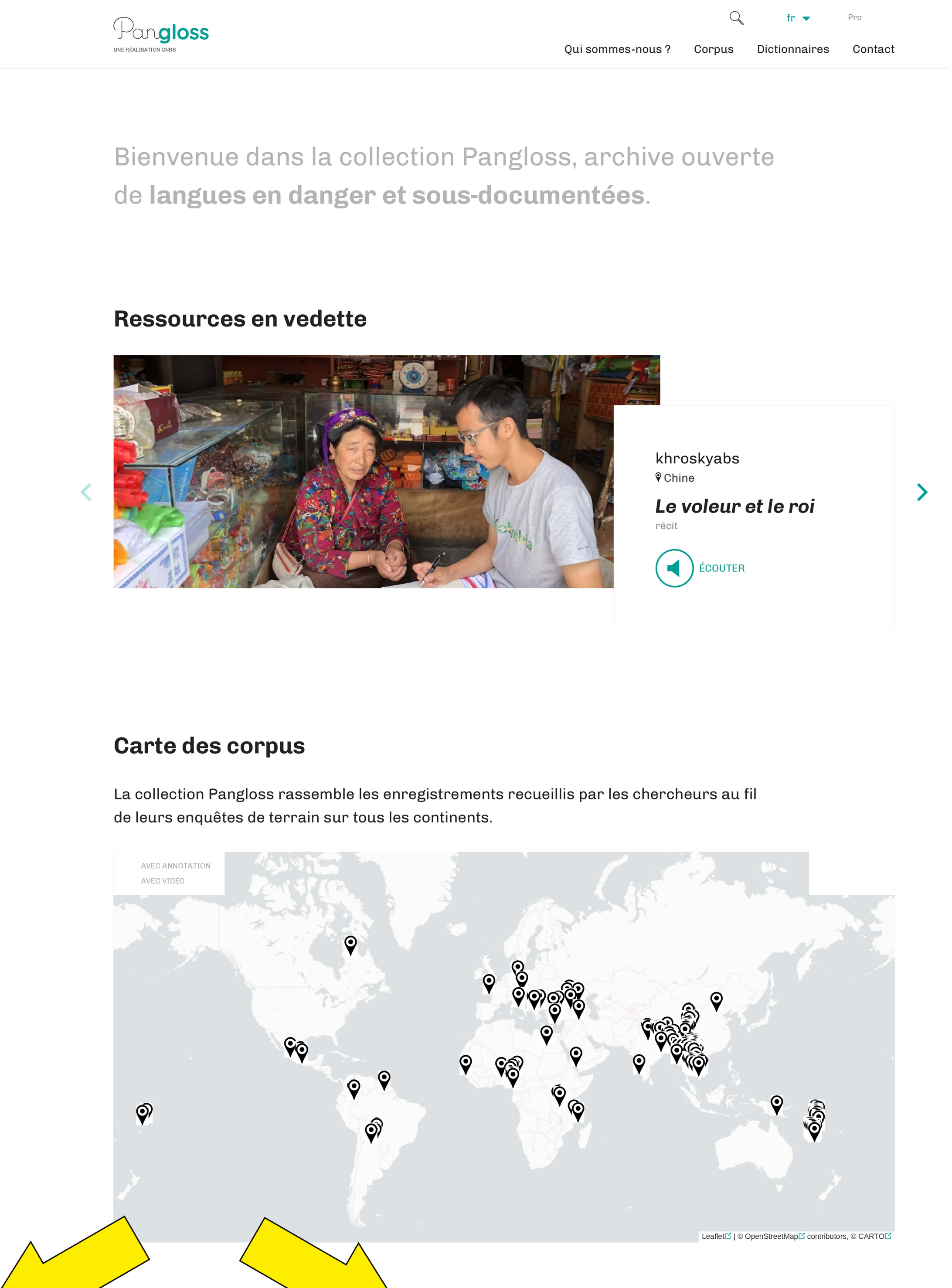
Facilitating NLP specialists’ access to language archive materials : an update

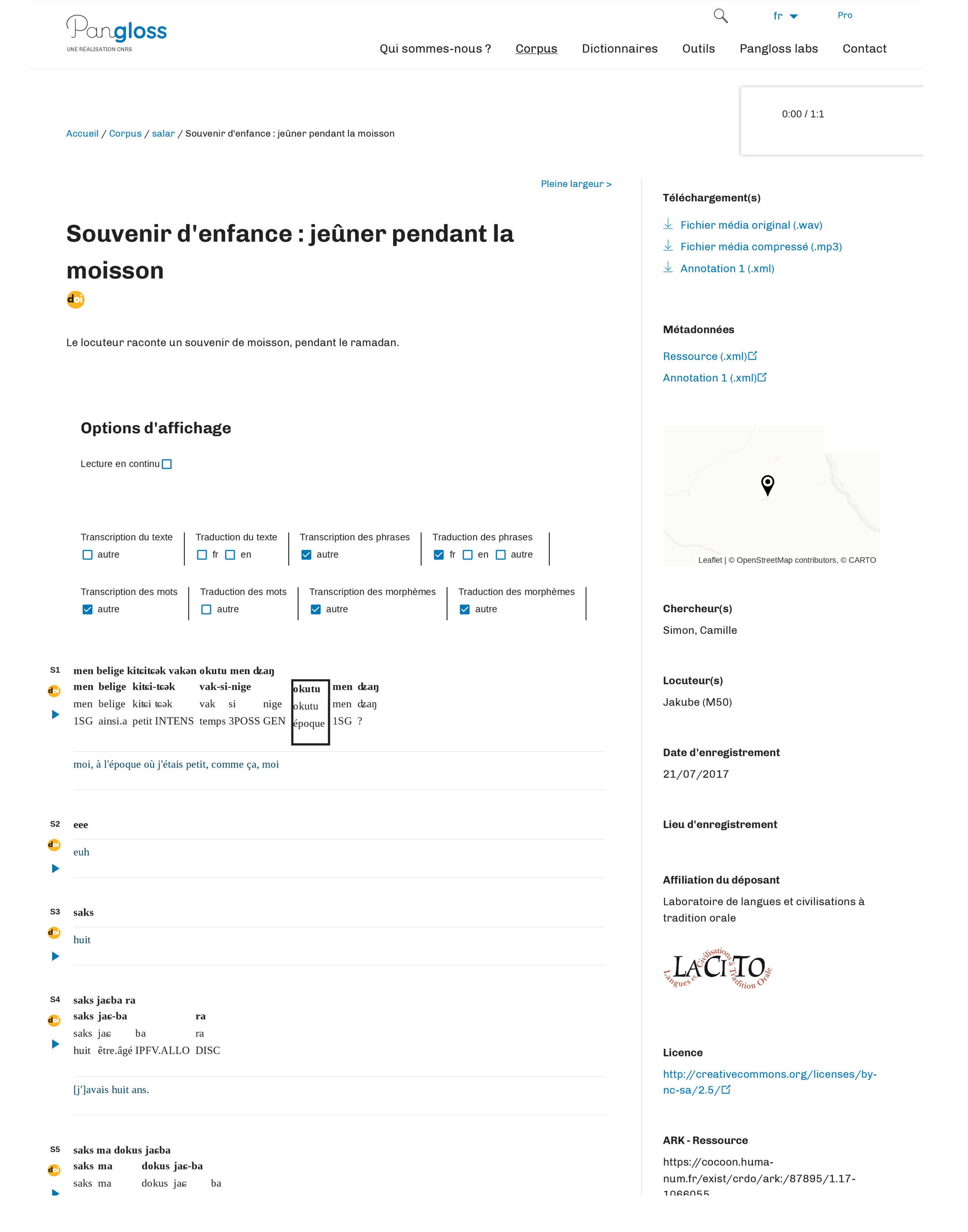
We present a software tool to assemble a great range of diverse datasets from the Pangloss collection (a multimedia open archive of under-documented languages). The tool ensures the reproducibility of experiments conducted on these data. As an example, two transcribed audio corpora of Chinese minority 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.

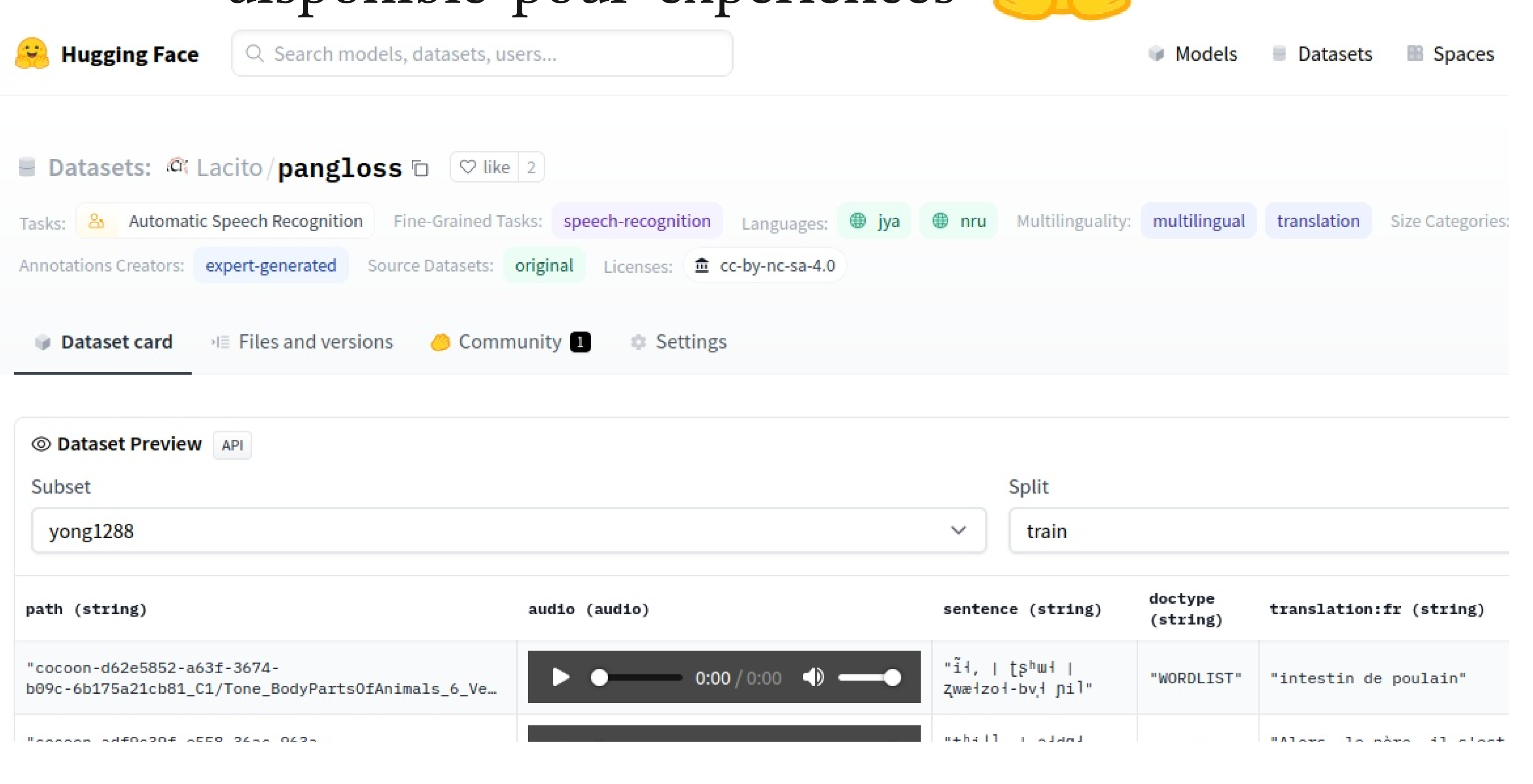
Difficulté identifiée : accès des TAListes aux corpus de langues rares. Archive ouverte Cocoon (Collection de Corpus Oraux Numériques) : pour consultation en ligne plutôt que téléchargement en masse.



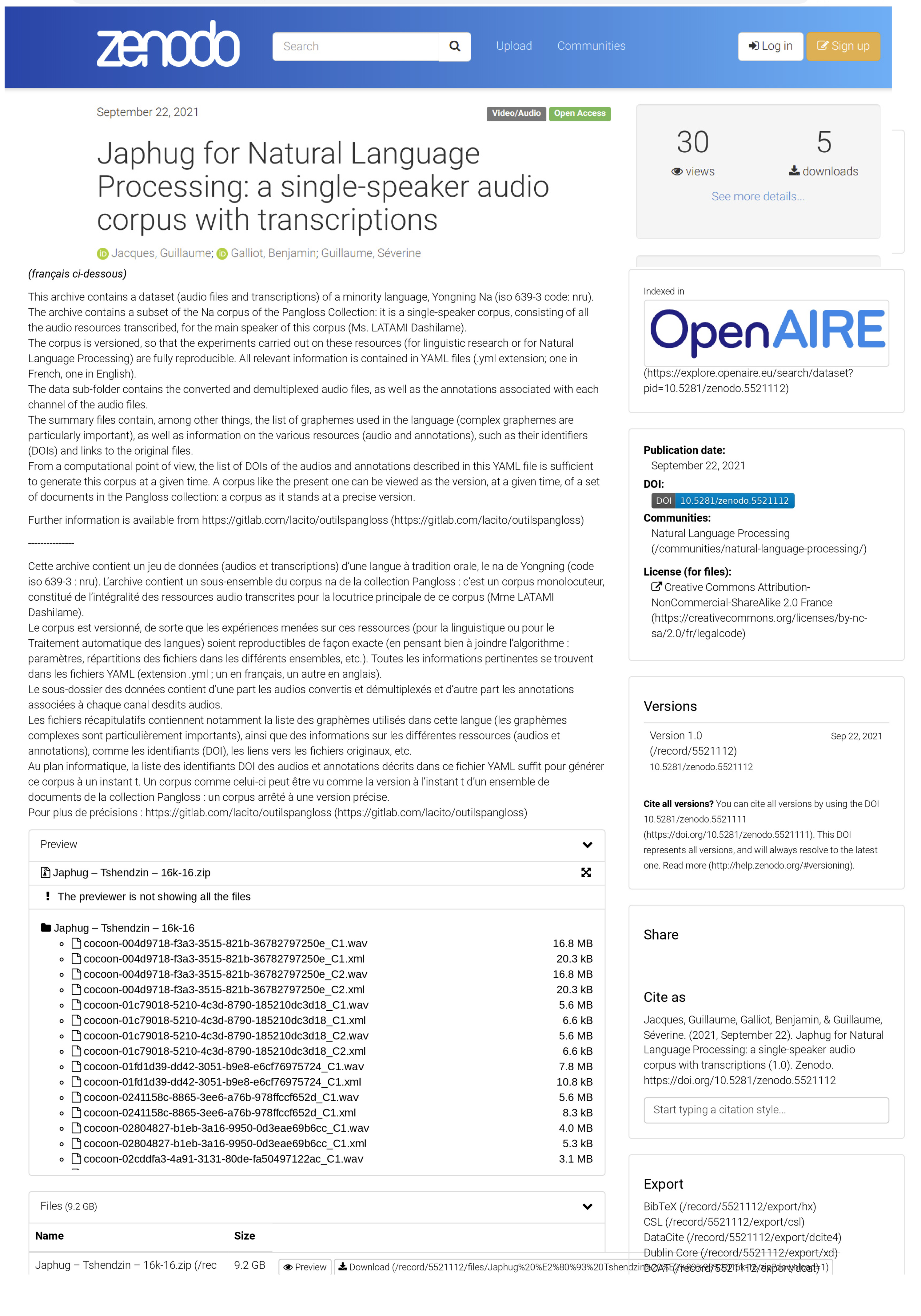


Besoin de reproductibilité des expériences.
Exemple d'utilisation : explorations en Reconnaissance Automatique de la Parole.

Hugging Face Datasets : disponible pour expériences



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



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 les 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) ;