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.

Besoin de reproductibilité des expériences.

Exemple d'utilisation : explorations en Reconnaissance Automatique de la Parole.
Voir exposé oral de Macaire et al.

macairececile / **internship_lacito_2021** Public

Code Issues Pull requests Actions Projects Wiki Security Insights

main

internship_lacito_2021 / src / xlsr / models / generate_predictions.py / Jump to

macairececile update

1 contributor

180 lines (151 sloc) | 7.75 KB

```
1  # -*- coding: utf-8 -*-
2  # ----------- Libraries ----------- #
3  import argparse
4  import difflib
5  import itertools
6  from os import listdir
7  from os.path import isfile, join
8
9  import librosa
10 import torchaudio
11 from datasets import load_dataset
12 import pandas as pd
```

Outil générique pour création de jeux de données **intégralement documentés** : Outils-Pangloss

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.

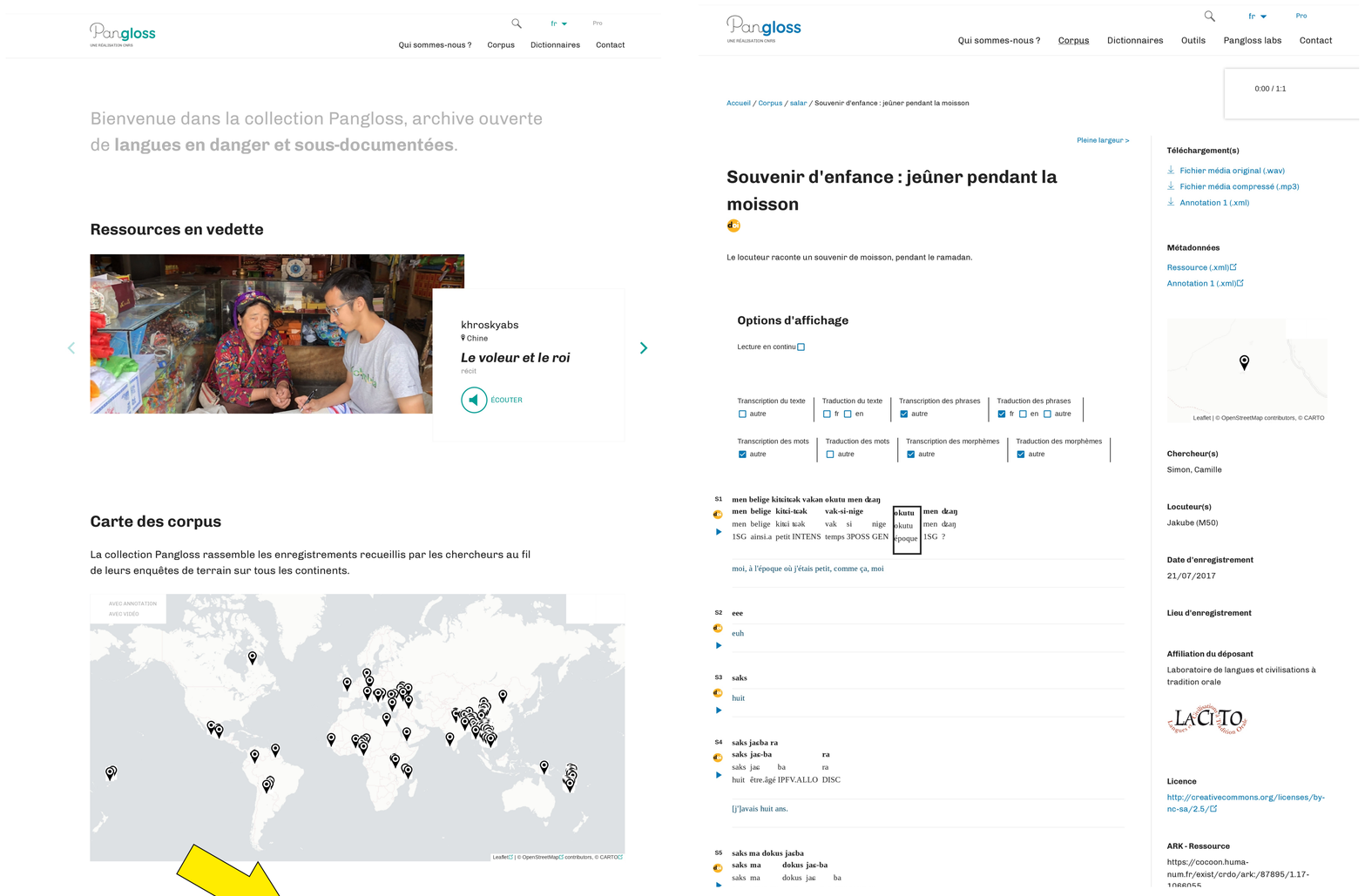
Utilisation

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`) ;

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



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