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, Guillaume Wisniewski, Séverine Guillaume, Guillaume Jacques, Alexis Michaud

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.

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 Anastasopoulos et al., 2020. 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. Cet exposé, qui fait suite à une précédente communication sur le même thème (Galliot et al., 2021), présente notre contribution à cette thématique. Suivant l'exemple de la publication du corpus mbochi (bantou) (Godard et al., 2018), nous avons déposé dans Zenodo et dans Hugging Face Datasets deux corpus audio (avec transcriptions) de langues rares : le japhug et le na, langues minoritaires de Chine. Ces corpus ont été utilisés dans des travaux en reconnaissance automatique de la parole (Adams et al., 2018; Adams et al., 2021; Guillaume, Wisniewski, Macaire, et al., 2022; Macaire, 2021) et dans des réflexions interdisciplinaires associant TAListes et linguistes (Guillaume, Wisniewski, Galliot, et al., 2022; Michaud et al., 2018 ; Michaud et al., 2020). Les corpus sont disponibles en ligne dans une archive ouverte, la collection Pangloss [1] (Michaud et al., 2016), archive ouverte de langues rares hébergée par la plateforme Cocoon [2], de sorte que l'accès en est ouvert.

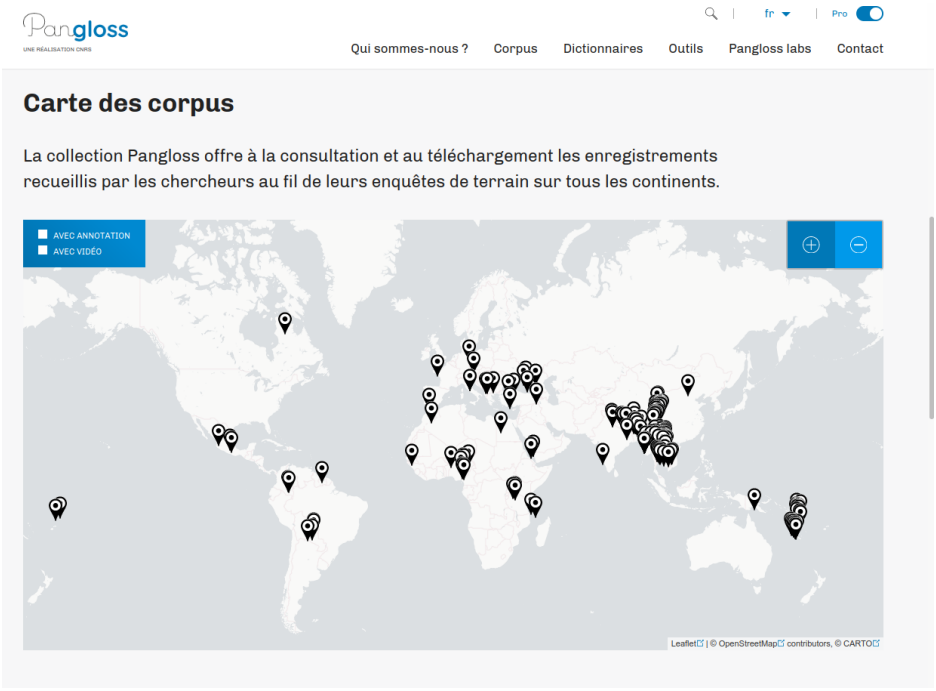


Figure 1 – Pangloss – carte des langues

Néanmoins, les transcriptions de ces corpus sont enrichies et revues au fil des années, et de nouveaux documents s'y ajoutent, de sorte que renvoyer à l'archive ouverte elle-même ne constitue pas une

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

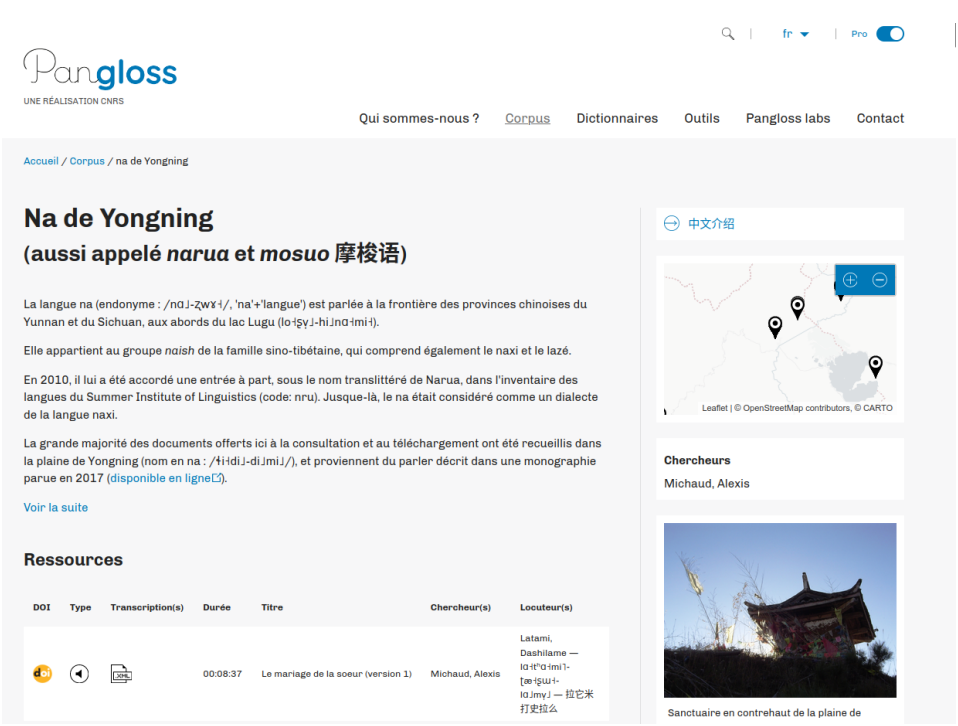


FIGURE 2 – Pangloss – page d'un corpus : introduction, et liste des ressources

référence suffisamment précise pour parvenir à la reproductibilité des expériences 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 d'un état donné de ces deux corpus, ainsi rendu accessibles en quelques clics, sous une forme stabilisée, dans Zenodo (§2) puis parmi les jeux de données Hugging Face (§3). Mais l'avancée la plus importante à nos yeux est la création d'un script, OutilPangloss (§4), pour panacher à son aise parmi les corpus de la collection Pangloss tout en conservant une garantie de reproductibilité (grâce au système de versionnage dont bénéficient les documents déposés en archive).

2 Les dépôts Zenodo : liens d'accès et choix techniques

Lieu de dépôt Les deux corpus na et japhug 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.5521112` pour le japhug, et `10.5281/zenodo.5336698` pour le na.

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. La logique qui préside à la constitution de ces corpus versionnés pour expériences de TAL s'éloigne de celle 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 exemple est fourni en Figure 3.

(a) données (b) métadonnées

FIGURE 3 – Fichiers XML des ressources d'annotations

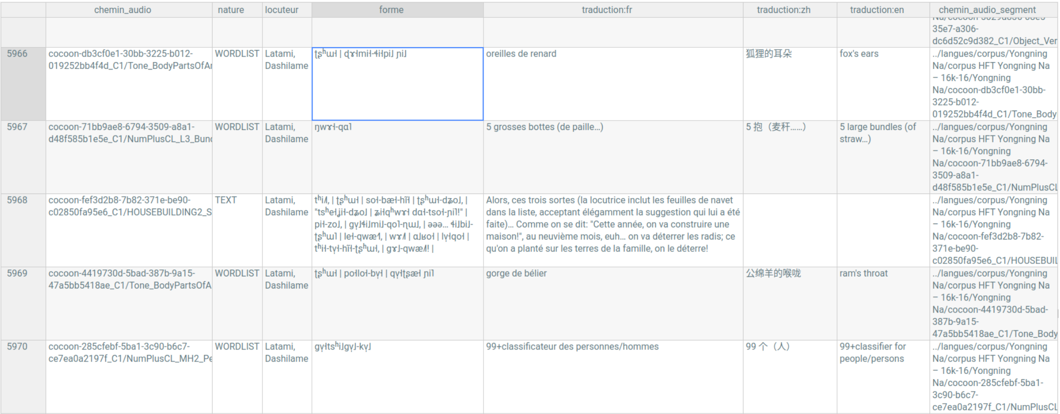
Un prétraitement élémentaire a été effectué, afin de ne pas imposer aux utilisateurs de devoir prendre connaissance d'un certain nombre de conventions choisies par les déposants. En particulier, lors de la transcription de textes, il arrive que des retouches soient apportées, qui éloignent la transcription 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 Les corpus au format Hugging Face

Les jeux de données Hugging Face (HF) constituent actuellement un lieu central pour les TAListes à la recherche de corpus aisément utilisables.Le formatage des corpus au format HF est donc apparu opportun dans la perspective qui consiste à porter des données de langues rares à l'attention de chercheuses et chercheurs en TAL.

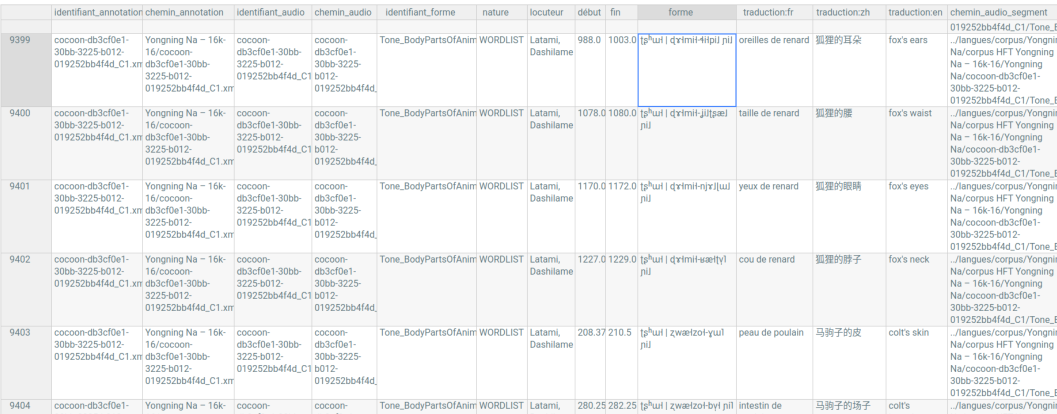
Préparation de l'archive formatée Chaque fichier d'annotations est divisé en autant de fichiers qu'il comporte d'unités de premier niveau (phrases pour les textes, mots pour les listes de vocabulaire). Ces fichiers portent le nom de ces unités, et sont placés dans des dossiers du nom de la ressource

d'origine. Ces données, structurées de manière tabulaire, sont partitionnées en trois jeux (entraînement, validation et test), donc trois fichiers CSV.



	chemin_audio	nature	locuteur	forme	traduction:fr	traduction:zh	traduction:en	chemin_audio_segment
								35e7-a306-dc6d52c9d382_C1/Object_Ver
5966	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1/Tone_BodyPartsOfAr	WORDLIST	Latami, Dashilame	ʈʂʰɯ˧ | ɖɤ˧mi˧-ɬi˧pi˩ ɲi˩	oreilles de renard	狐狸的耳朵	fox's ears	../langues/corpus/Yongning Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1/Tone_Body
5967	cocoon-71bb9ae8-6794-3509-a8a1-d48f585b1e5e_C1/NumPlusCL_L3_Bun	WORDLIST	Latami, Dashilame	ŋwɤ˧-qɑ˥	5 grosses bottes (de paille...)	5 抱（麦秆......）	5 large bundles (of straw...)	../langues/corpus/Yongning Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-71bb9ae8-6794-3509-a8a1-d48f585b1e5e_C1/NumPlusCL
5968	cocoon-fef3d2b8-7b82-371e-be90-c02850fa95e6_C1/HOUSEBUILDING2_S	TEXT	Latami, Dashilame	tʰi˩˥, | ʈʂʰɯ˧ | so˧-bæ˧-hĩ˧ | ʈʂʰɯ˧-dʑo˩, | "tsʰe˧ʝi˧-dʑo˩ | ʑi˧qʰwɤ˧ dɑ˧-tso˧-ɲi˥!" | pi˧-zo˩, | gv̩˩ɬi˩mi˩-qo˥-ŋɯ˩, | əəə... ɬi˩bi˩-ʈʂʰɯ˥ | le˧-qwæ˦, | wɤ˩˥ | ɑʁo˧ | lv̩˧qo˧ | tʰi˧-tv̩˧-hĩ˧-ʈʂʰɯ˧, | gɤ˩-qwæ˩˥! |	Alors, ces trois sortes (la locutrice inclut les feuilles de navet dans la liste, acceptant élégamment la suggestion qui lui a été faite)... Comme on se dit: "Cette année, on va construire une maison!", au neuvième mois, euh... on va déterrer les radis; ce qu'on a planté sur les terres de la famille, on le déterre!			../langues/corpus/Yongning Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-fef3d2b8-7b82-371e-be90-c02850fa95e6_C1/HOUSEBUIL
5969	cocoon-4419730d-5bad-387b-9a15-47a5bb5418ae_C1/Tone_BodyPartsOfA	WORDLIST	Latami, Dashilame	ʈʂʰɯ˧ | po˧lo˧-bv̩˧ | qv̩˧ʈʂæ˧ ɲi˥	gorge de bélier	公绵羊的喉咙	ram's throat	../langues/corpus/Yongning Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-4419730d-5bad-387b-9a15-47a5bb5418ae_C1/Tone_Body
5970	cocoon-285cfebf-5ba1-3c90-b6c7-ce7ea0a2197f_C1/NumPlusCL_MH2_Pe	WORDLIST	Latami, Dashilame	gv̩˧tsʰi˩gv̩˩-kv̩˩	99+classificateur des personnes/hommes	99 个（人）	99+classifier for people/persons	../langues/corpus/Yongning Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-285cfebf-5ba1-3c90-b6c7-ce7ea0a2197f_C1/NumPlusCL

(a) données pour le format Hugging Face



	identifiant_annotation	chemin_annotation	identifiant_audio	chemin_audio	identifiant_forme	nature	locuteur	début	fin	forme	traduction:fr	traduction:zh	traduction:en	chemin_audio_segment
														019252bb4f4d_C1/Tone_E
9399	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1.xm	Yongning Na – 16k-16/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1.xm	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_	Tone_BodyPartsOfAnim	WORDLIST	Latami, Dashilame	988.0	1003.0	ʈʂʰɯ˧ | ɖɤ˧mi˧-ɬi˧pi˩ ɲi˩	oreilles de renard	狐狸的耳朵	fox's ears	../langues/corpus/Yongnir Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1/Tone_E
9400	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1.xm	Yongning Na – 16k-16/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1.xm	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_	Tone_BodyPartsOfAnim	WORDLIST	Latami, Dashilame	1078.0	1080.0	ʈʂʰɯ˧ | ɖɤ˧mi˧-ʝi˩ʈʂæ˩ ɲi˩	taille de renard	狐狸的腰	fox's waist	../langues/corpus/Yongnir Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1/Tone_E
9401	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1.xm	Yongning Na – 16k-16/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1.xm	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_	Tone_BodyPartsOfAnim	WORDLIST	Latami, Dashilame	1170.0	1172.0	ʈʂʰɯ˧ | ɖɤ˧mi˧-njɤ˩ʈɯ˩ ɲi˩	yeux de renard	狐狸的眼睛	fox's eyes	../langues/corpus/Yongnir Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1/Tone_E
9402	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1.xm	Yongning Na – 16k-16/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1.xm	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_	Tone_BodyPartsOfAnim	WORDLIST	Latami, Dashilame	1227.0	1229.0	ʈʂʰɯ˧ | ɖɤ˧mi˧-ʁæ˧ʈv̩˥ ɲi˩	cou de renard	狐狸的脖子	fox's neck	../langues/corpus/Yongnir Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1/Tone_E
9403	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1.xm	Yongning Na – 16k-16/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1.xm	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1	cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_	Tone_BodyPartsOfAnim	WORDLIST	Latami, Dashilame	208.37	210.5	ʈʂʰɯ˧ | ʐwæ˧zo˧-ɣɯ˥ ɲi˩	peau de poulain	马驹子的皮	colt's skin	../langues/corpus/Yongnir Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1/Tone_E
9404	cocoon-db3cf0e1-	Yongning Na – 16k-	cocoon-	cocoon-	Tone_BodyPartsOfAnim	WORDLIST	Latami,	280.25	282.25	ʈʂʰɯ˧ | ʐwæ˧zo˧-bv̩˧ ɲi˩	intestin de	马驹子的场子	colt's	../langues/corpus/Yongnir

(b) données au format Hugging Face (après partitionnement aléatoire)

FIGURE 4 – OutilsPangloss : fichiers CSV

Certaines métadonnées pertinentes, comme la nature de la ressource (liste de mots ou texte), apparaissent également dans le format tabulaire. Les données sont ensuite aléatoirement partitionnées (par défaut au niveau des phrases, mais il est possible de choisir le niveau des fichiers) en trois jeux (entraînement, validation et test), donc trois fichiers CSV. Enfin, toutes ces données (fichiers audios des phrases et 3 fichiers CSV) sont téléversées sur un serveur.

Préparation du jeu de données public La seconde partie consiste à préparer formellement les nouveaux jeux de données [3]. Pour ce faire, il faut préparer une description précise de chaque corpus. Une attention particulière est portée à l'encodage de la langue concernée [4], point délicat dans le cas des langues pas ou peu standardisées qui constituent le coeur de métier de la collection Pangloss. Un

3. Disponibles ici : https://huggingface.co/datasets/Lacito/pangloss
4. https://github.com/huggingface/datasets/issues/4881

script [5] Python automatise la création du jeu de données depuis les archives des corpus préalablement téléversées et accessibles. Ce script, de complexité variable selon les jeux de données, sert notamment à préciser où sont stockées les archives formatées, les types de données de chaque colonne et leurs noms correspondants. Il est ensuite possible de visualiser en ligne directement les données des corpus et d'écouter les précieux segments audio : voir Figure 5.

Il a été choisi de créer un jeu de données global du nom de Pangloss, dont les diverses langues (japhug [6], na de Yongning [7]) constituent un sous-jeu.

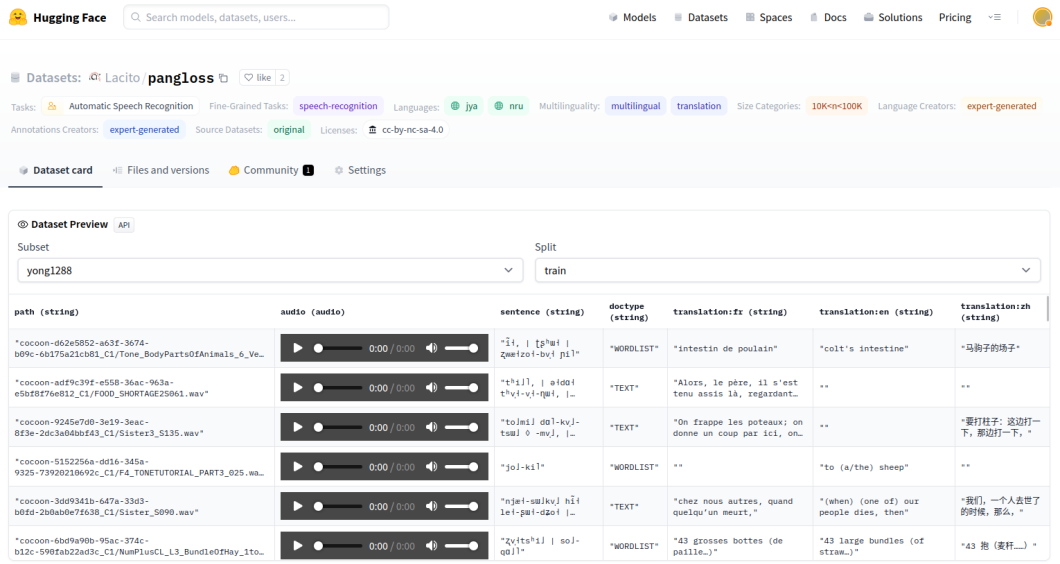


FIGURE 5 – Hugging Face : page des corpus Pangloss

Il est dès lors possible en deux lignes [8] de charger le jeu de données voulu et de commencer à les utiliser.

4 Un outil pour constituer de nouveaux jeux de données

L'outil élaboré à l'occasion de la préparation des corpus déposés dans Zenodo et Hugging Face, intitulé OutilsPangloss [9], 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 (telle qu'elle figure sur Pangloss). Elle peut également fournir une liste d'expressions rationnelles de modifications si des traitements

5. https://huggingface.co/datasets/Lacito/pangloss/blob/main/pangloss.py
6. https://huggingface.co/datasets/Lacito/pangloss/viewer/japh1234/train
7. https://huggingface.co/datasets/Lacito/pangloss/viewer/yong1288/train
8. `import datasets` puis
`datasets.load_dataset("pangloss", "japh1234")` pour le japhug et
`datasets.load_dataset("pangloss", "yong1288")` pour le na de Yongning.
9. https://gitlab.com/lacito/outilspangloss

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` : voir Figure 6. 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.

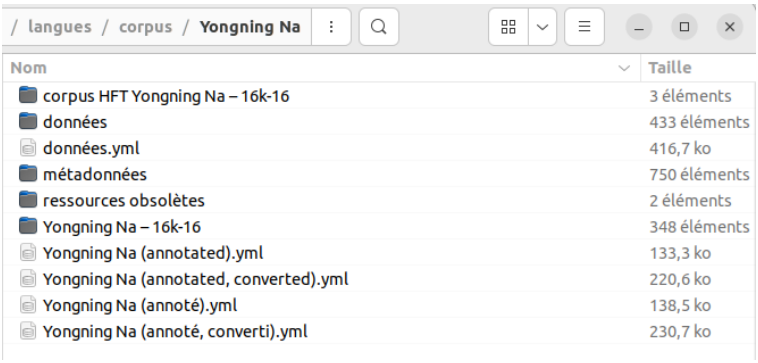
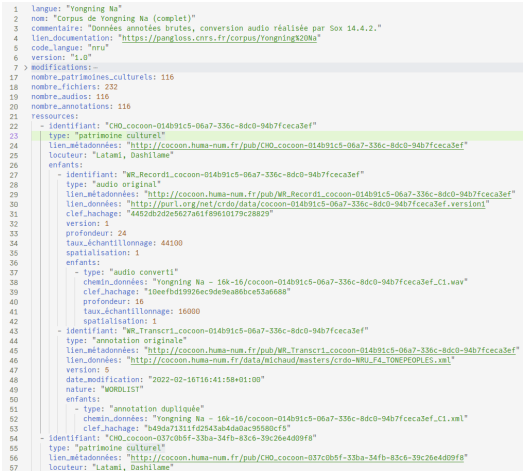


FIGURE 6 – OutilsPangloss : dossier cible local d'un corpus

```yaml
langue: Yongning Na
modifications:…
corpus:
  chemin: "../langues/corpus/Yongning Na"
  nom complet: "Corpus de Na de Yongning complet"
  commentaire: "Données complètes."
  langue: fr
  sous-corpus:
  - récapitulatifs:
    - nom de fichier: "Na de Yongning (annoté)"
      nom complet: "Corpus de Na de Yongning (complet)"
      commentaire: "Données annotées brutes."
      langue: fr
    - nom de fichier: "Yongning Na (annotated)"
      nom complet: "Yongning Na corpus (complete)"
      commentaire: "Raw annotated data."
      langue: en
  - sous-chemin: "Yongning Na - 16k-16"
    traitements:
      audio:
        taux d'échantillonnage: 16000
        profondeur: 16
        spatialisation: "séparer"
    récapitulatifs:
    - nom de fichier: "Na de Yongning (annoté, converti)"
      nom complet: "Corpus de Na de Yongning (complet)"
      commentaire: "Données annotées brutes, conversion audio réalisée par Sox 14.4.2."
      langue: fr
    - nom de fichier: "Yongning Na (annotated, converted)"
      nom complet: "Yongning Na corpus (complete)"
      commentaire: "Raw annotated data, audio conversion performed by Sox 14.4.2."
      langue: en
    conversions:
    - format: "HuggingFace-Transformers"
      sous-chemin: "corpus HFT Yongning Na - 16k-16"
      nom de fichier: "Na de Yongning (annoté, converti)"
      nom du corpus: "yong1288"
```

(a) configuration d'un corpus



```yaml
langue: "Yongning Na"
nom: "Corpus de Yongning Na (complet)"
commentaire: "Données annotées brutes, conversion audio réalisée par Sox 14.4.2."
lien_documentation: "https://pangloss.cnrs.fr/corpus/Yongning%20Na"
code_langue: "nru"
version: "1.0"
modifications:…
nombre_patrimoines_culturels: 116
nombre_fichiers: 232
nombre_audios: 116
nombre_annotations: 116
ressources:
  - identifiant: "CHO_cocoon-014b91c5-06a7-336c-8dc0-94b7fceca3ef"
    type: "patrimoine culturel"
    lien_métadonnées: "http://cocoon.huma-num.fr/pub/CHO_cocoon-014b91c5-06a7-336c-8dc0-94b7fceca3ef"
    locuteur: "Latami, Dashilome"
    enfants:
      - identifiant: "WR_Record1_cocoon-014b91c5-06a7-336c-8dc0-94b7fceca3ef"
        type: "audio original"
        lien_métadonnées: "http://cocoon.huma-num.fr/pub/WR_Record1_cocoon-014b91c5-06a7-336c-8dc0-94b7fceca3ef"
        lien_données: "http://purl.org/net/crdo/data/cocoon-014b91c5-06a7-336c-8dc0-94b7fceca3ef.version1"
        clef_hachage: "4452db2d2e5627a61f89610179c28829"
        version: 1
        profondeur: 24
        taux_échantillonnage: 44100
        spatialisation: 1
        enfants:
          - type: "audio converti"
            chemin_données: "Yongning Na - 16k-16/cocoon-014b91c5-06a7-336c-8dc0-94b7fceca3ef_C1.wav"
            clef_hachage: "10eefbd19926ec9de9ea86bce53a6688"
            profondeur: 16
            taux_échantillonnage: 16000
            spatialisation: 1
      - identifiant: "WR_Transcr1_cocoon-014b91c5-06a7-336c-8dc0-94b7fceca3ef"
        type: "annotation originale"
        lien_métadonnées: "http://cocoon.huma-num.fr/pub/WR_Transcr1_cocoon-014b91c5-06a7-336c-8dc0-94b7fceca3ef"
        lien_données: "http://cocoon.huma-num.fr/data/michaud/masters/crdo-NRU_F4_TONEPEOPLES.xml"
        version: 5
        date_modification: "2022-02-16T16:41:58+01:00"
        nature: "WORDLIST"
        enfants:
          - type: "annotation dupliquée"
            chemin_données: "Yongning Na - 16k-16/cocoon-014b91c5-06a7-336c-8dc0-94b7fceca3ef_C1.xml"
            clef_hachage: "b49da71311fd2543ab4da0ac95580cf5"
  - identifiant: "CHO_cocoon-037c0b5f-33ba-34fb-83c6-39c26e4d09f8"
    type: "patrimoine culturel"
    lien_métadonnées: "http://cocoon.huma-num.fr/pub/CHO_cocoon-037c0b5f-33ba-34fb-83c6-39c26e4d09f8"
    locuteur: "Latami, Dashilome"
```

(b) corpus versionné (liste des ressources, métadonnées...)

FIGURE 7 – OutilsPangloss : fichiers YML

5 Perspectives : enjeux d'une description exhaustive des jeux de données sans copie *en dur*

La transition en cours vers la Science ouverte porte une exigence fondamentale de reproductibilité des expériences, et le domaine des sciences de la parole et du Traitement Automatique des Langues ne fait pas exception (Garellek et al., 2020). 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).

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., & Michaud, A. (2018). Evaluating phonemic transcription of low-resource tonal languages for language documentation. *Proceedings of LREC 2018 (Language Resources and Evaluation Conference)*, 3356–3365. https://halshs.archives-ouvertes.fr/halshs-01709648

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., & Hill, N. (2021). User-friendly automatic transcription of low-resource languages: Plugging ESPnet into Elpis. *Proceedings of ComputEL-4: Fourth Workshop on the Use of Computational Methods in the Study of Endangered Languages*. https://halshs.archives-ouvertes.fr/halshs-03030529

Anastasopoulos, A., Cox, C., Neubig, G., & Cruz, H. (2020). Endangered languages meet Modern NLP. *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, 39–45. https://doi.org/10.18653/v1/2020.coling-tutorials.7

Galliot, B., Wisniewski, G., Guillaume, S., Michaud, A., Rossato, S., Nguyên, M.-C., & Fily, M. (2021). Deux corpus audio transcrits de langues rares (japhug et na) normalisés en vue d'expériences en traitement du signal. *Journées scientifiques du Groupement de recherche "Linguistique informatique, formelle et de terrain" (GDR LIFT)*. https://halshs.archives-ouvertes.fr/halshs-03475436

Garellek, M., Gordon, M., Kirby, J., Lee, W.-S., Michaud, A., Mooshammer, C., Niebuhr, O., Recasens, D., Roettger, T. B., Simpson, A., et al. (2020). Toward open data policies in phonetics: What we can gain and how we can avoid pitfalls. *Journal of Speech Science*, *9*(1), 3–16.

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

Guillaume, S., Wisniewski, G., Galliot, B., Nguyễn, M.-C., Fily, M., Jacques, G., & Michaud, A. (2022). Plugging a neural phoneme recognizer into a simple language model: a workflow for low-resource settings. *Interspeech 2022 - 23rd Annual Conference of the International Speech Communication Association*. https://doi.org/10.5281/zenodo.5521111

Guillaume, S., Wisniewski, G., Macaire, C., Jacques, G., Michaud, A., Galliot, B., Coavoux, M., Rossato, S., Nguyễn, M.-C., & Fily, M. (2022). Fine-tuning pre-trained models for Automatic Speech Recognition: experiments on a fieldwork corpus of Japhug (Trans-Himalayan family). *ComputEL-5 5th Workshop on Computational Methods for Endangered Languages (ComputEL-5)*. https://halshs.archives-ouvertes.fr/halshs-03647315

Macaire, C. (2021). *Recognizing lexical units in low-resource language contexts with supervised and unsupervised neural networks* (Research Report). LACITO (UMR 7107). https://hal.archives-ouvertes.fr/hal-03429051

Michaud, A., Adams, O., Cohn, T., Neubig, G., & 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. http://hdl.handle.net/10125/24793

Michaud, A., Adams, O., Cox, C., Guillaume, S., Wisniewski, G., & 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). https://halshs.archives-ouvertes.fr/halshs-02881731/

Michaud, A., Guillaume, S., Jacques, G., Mac, D.-K., Jacobson, M., Pham, T.-H., & Deo, M. (2016). Contribuer au progrès solidaire des recherches et de la documentation : la Collection Pangloss et la Collection AuCo. *Journées d'Etude de la Parole 2016*, *1*, 155-163. https://halshs.archives-ouvertes.fr/halshs-01341631

Vasile, A., Guillaume, S., Aouini, M., & 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. https://halshs.archives-ouvertes.fr/halshs-02870206