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

Benjamin Galliot, Guillaume Wisniewski, Séverine Guillaume, Guillaume Jacques, Alexis Michaud

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

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

Abstract

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.

Résumé

Faciliter l'accès des praticiens du traitement automatique des langues à des jeux de données de langues rares : un deuxième point d'étape

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.

Keywords: benchmark datasets, computational language documentation, endangered languages.

Mots-clés : corpus de référence, documentation computationnelle des langues, langues rares.

1 Introduction

The deployment of automatic speech processing tools clearly has important implications for language documentation, at a time when the decline in linguistic diversity is accelerating (Kik et al., 2021), in parallel with the decline in biodiversity. Conversely, less-documented languages present a whole range of challenges to computational research, the interest of which is increasingly clearly perceived (Anastasopoulos et al., 2020). In this context, providing easily accessible, clearly versioned and user-friendly corpora of less-studied languages appears as a central necessity. This presentation, which follows a previous paper on the same theme (Galliot et al., 2021), presents our contribution to this endeavour. Following the example of the publication of the Mbochi (Bantu) corpus (Godard et al., 2018), we have deposited in Zenodo and in Hugging Face Datasets audio corpora (with transcriptions) of Japhug and Na, two minority languages of China.

These corpora have been used in automatic speech recognition work (Adams et al., 2018; Adams et al., 2021; Guillaume, Wisniewski, Macaire, et al., 2022; Macaire, 2021) and in interdisciplinary reflections associating NLP specialists and linguists (Guillaume, Wisniewski, Galliot, et al., 2022; Michaud et al., 2018; Michaud et al., 2020). The corpora are available online in an open archive, the Pangloss Collection[1] (Michaud et al., 2016), an open archive of (mostly) endangered languages, itself hosted by the Cocoon data repository[2], ensuring availability in open access.

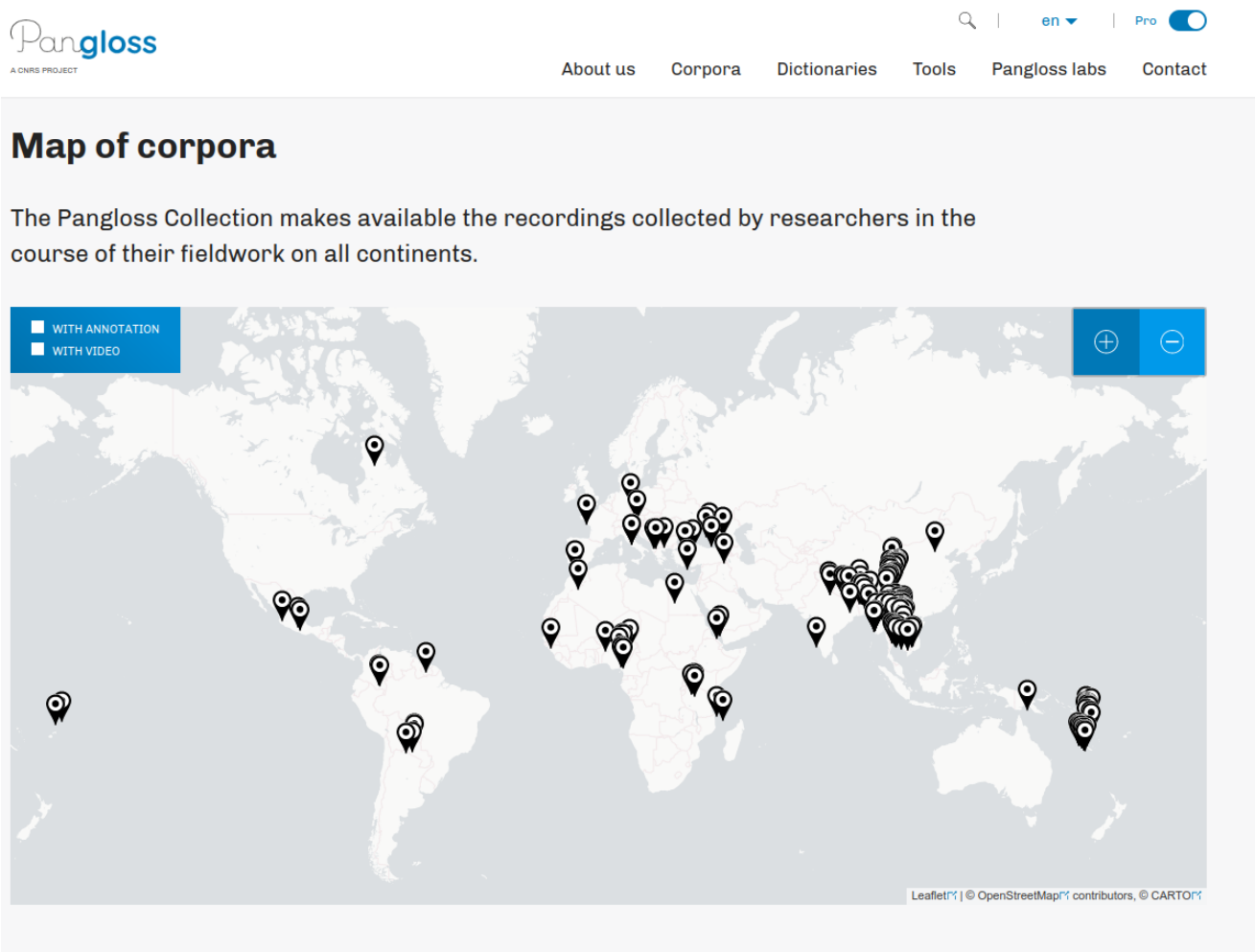


Figure 1: Pangloss – map of languages

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

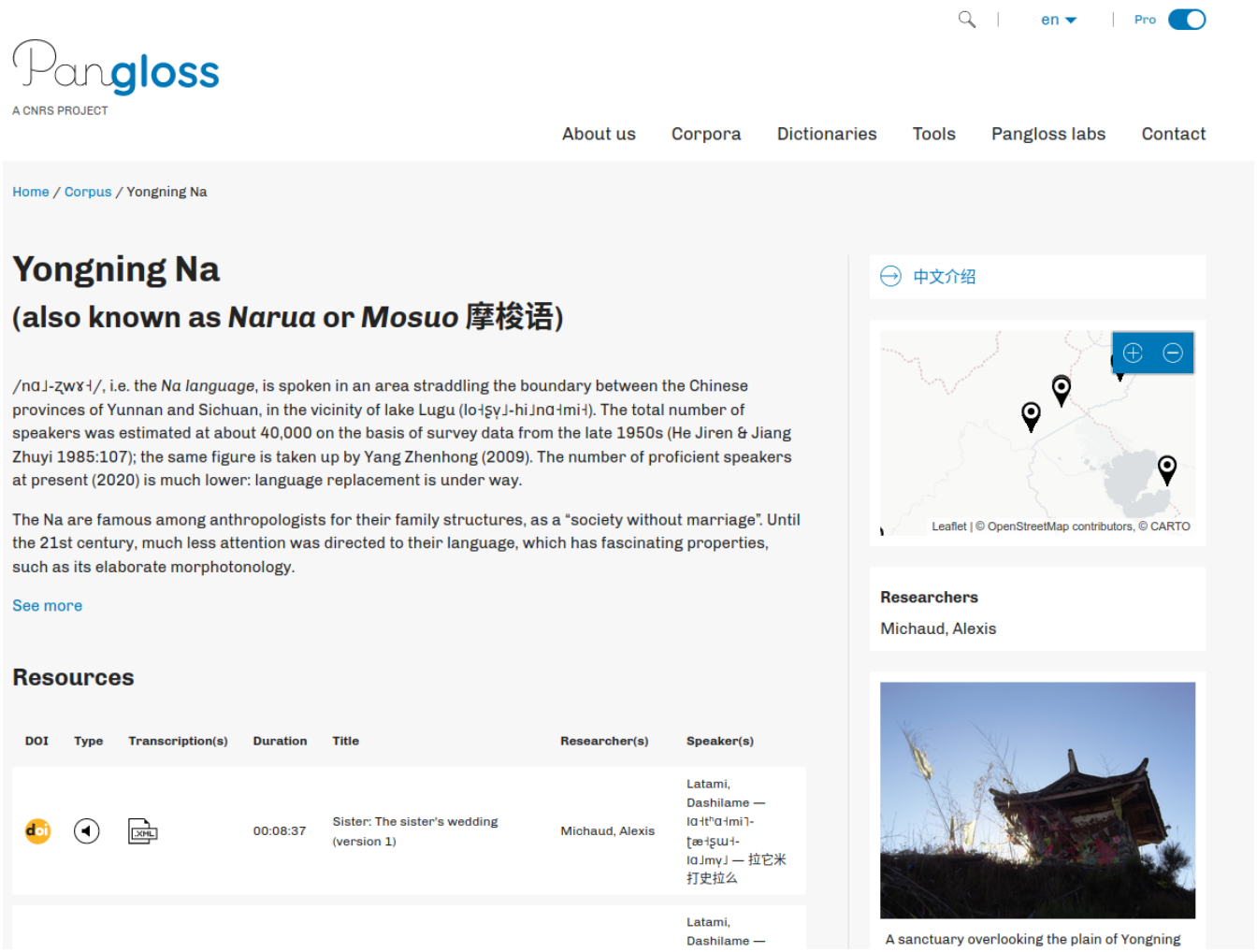


Figure 2: Pangloss – page of a corpus: introduction and list of available multimedia resources

The transcriptions of these corpora are enriched and revised over the years, however, and new documents are added, so that referring to the open archive itself is not a sufficiently precise reference to achieve actual reproducibility of the experiments conducted on these data. Even though the primary data (the audio and video files) do not change, the transcriptions constitute "lukewarm data", which get improved and modified slowly over time.

We have therefore made a deposit of a given state of these two corpora, making them accessible in a few clicks, in a stabilized form, in Zenodo (§3) and among the Hugging Face Datasets (§4). But the most important advance, in our view, is the creation of a script, OutilsPangloss (§2), to select among the corpora of the Pangloss Collection while maintaining a guarantee of full reproducibility (thanks to the versioning system of documents deposited in the archive).

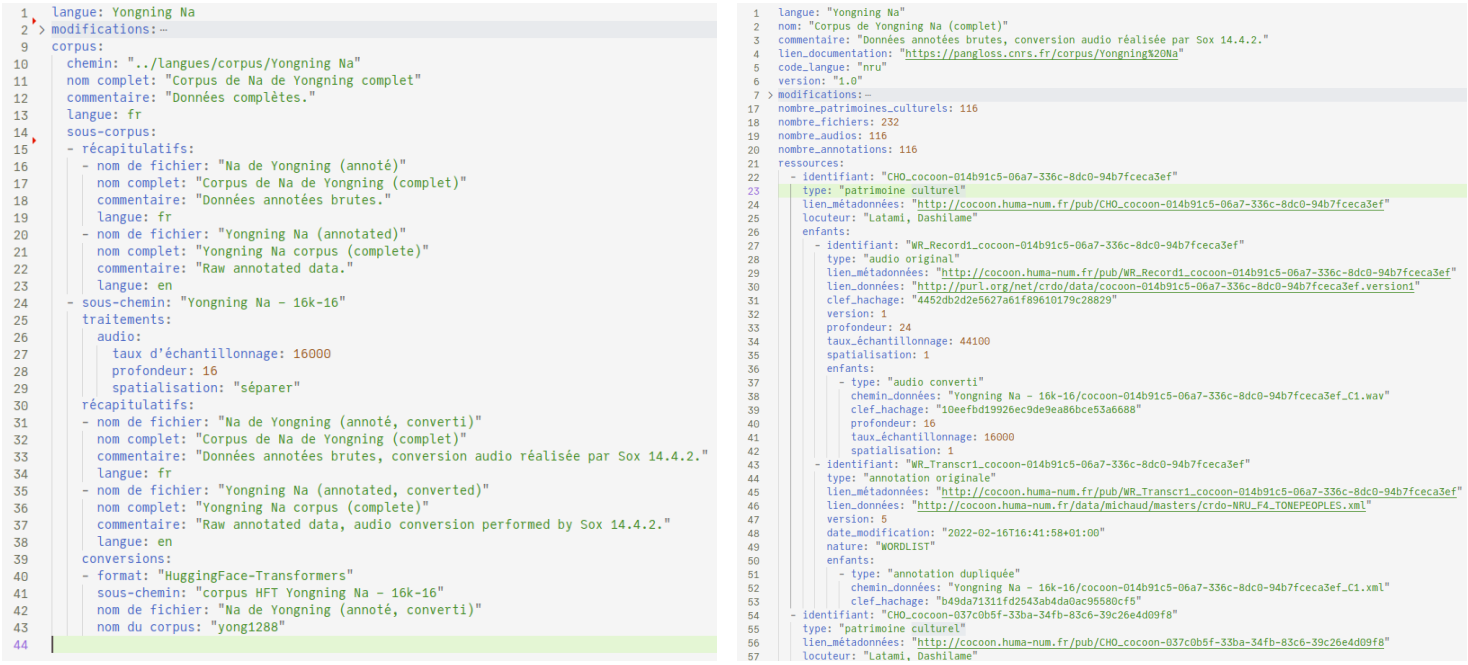
2 A tool to build new datasets

The tool developed during the preparation of the corpora deposited in Zenodo and Hugging Face, entitled OutilsPangloss[3], allows for putting together an open-ended range of new datasets. It consists of a toolbox in Julia language, which, among other functionalities, creates (sub)corpora of less-documented language data from the Pangloss Collection. The user fills in a YAML file (of which various examples are provided, see Figure 3a), specifying the names of the desired languages (as they

[3] https://gitlab.com/lacito/outilspangloss

appears in the Pangloss Collection). It can also provide a list of regular expressions for modifications in case processing of annotations is to be carried out (such as deletions or rearrangements of text blocks). It is possible to filter by speaker. Processing on the audio can also be set, to choose the sampling rate and bit-depth, and to separate the different tracks of multi-channel files into mono files (demultiplexing).

After harvesting (in Sparql), metadata checks (hashing, versions, etc.), data checks (mistake detection, etc.), and downloads, a general summary file (see Figure 3b) is generated in the target folder, alongside data and metadata folders (see Figure 4). The information contained in this summary file is sufficient to reproduce exactly, at any time, an experiment conducted with the dataset it describes.



```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) corpus configuration

```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, Dashilame"
    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: "b49da71311fd2543ab4da0ac9558Ocf5"
  - 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, Dashilame"
```

(b) Versioned corpus (list of resources, metadata...)

Figure 3: OutilsPangloss : YML files

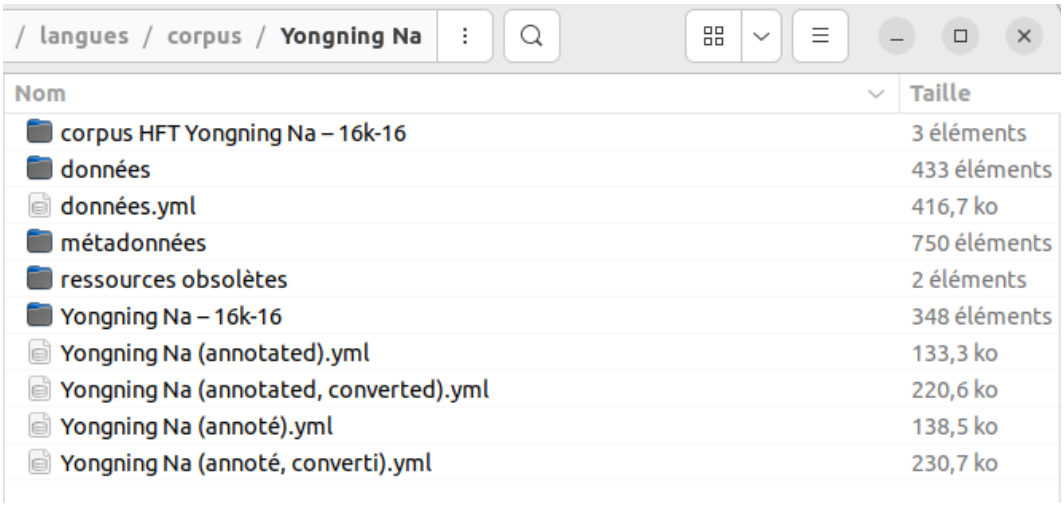
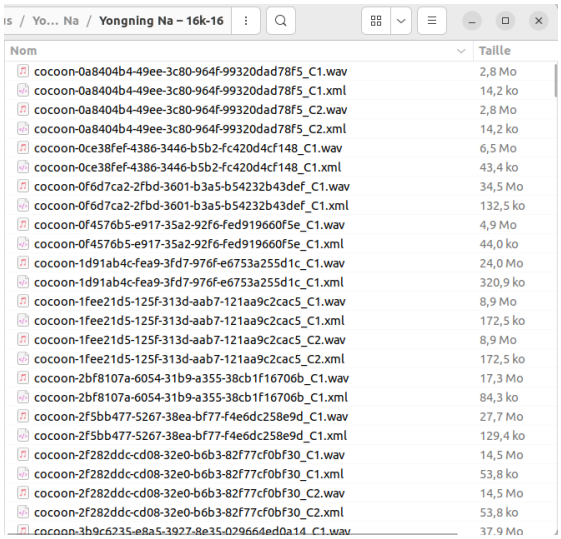
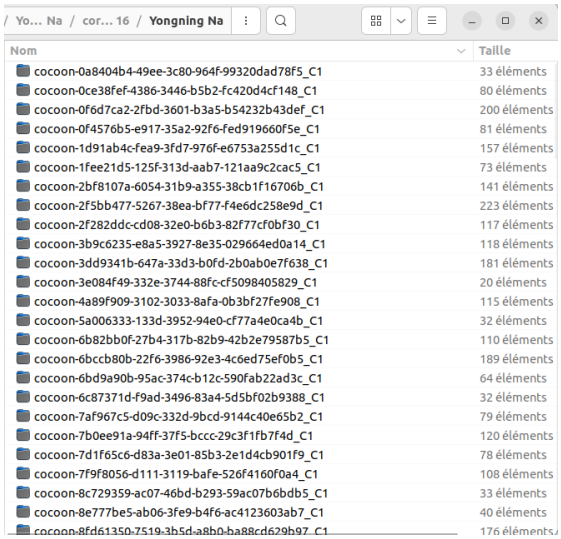


Figure 4: OutilsPangloss: local target folder of a corpus



(a) processed resources



(b) processed resources for the Hugging Face format

Figure 5: OutilsPangloss : données

3 The Zenodo deposits: access links and explanations about technical choices

Links to the repository The Na and Japhug corpora were uploaded to Zenodo, where they constitute deposits 5336698 (Na) and 5521112 (Japhug), respectively. An entire corpus is identified by a DOI (*digital object identifier*): `10.5281/zenodo.5521112` for the Japhug corpus, and `10.5281/zenodo.5336698` for the Na corpus.

The same type of identifier has been deployed for the Pangloss Collection, the open archive where the corpora are deposited, but with a completely different granularity: a DOI for each document (Vasile et al., 2020), a choice which is suitable for linguists who wish to reference data at a highly specific level (a text and, within a text, a specific utterance) but does not give a handle on an entire corpus.

Audio files Audio files were downgraded to 16-bit, 16 kHz, mono (see Figure 5a). The logic behind this choice of parameters is that NLP experiments have requirements that differ from those of long-term archiving in the Pangloss Collection. The size of the two datasets deposited in Zenodo is compatible (as of today) with experiments carried out on a laptop computer: 1.8 GB for Na, 9.2 GB for Japhug.

Annotations The annotations are in the original format: XML structured according to a simple hierarchy (a text is composed of sentences, themselves composed of words, themselves composed of morphemes). An example is provided in Figure 6.

Some basic preprocessing has been carried out, so that users do not have to learn about the various conventions chosen by the depositors, which vary across corpora. In particular, when transcribing texts, it is not infrequent for language consultants and language workers (typically linguists) to make decisions that create an edit distance between the transcription and what is actually said in the recording. A convention used in the Na corpus is that added passages are placed within square brackets. Another convention is that passages that the language consultants wish to see removed from the "smoothed" transcription are placed between angle brackets. At preprocessing, the passages within square brackets were deleted, and the angle brackets were removed, thereby ensuring the closest match between audio and transcriptions.

(a) data (b) metadata

Figure 6: XML files of annotation resources

4 The Pangloss corpus in Hugging Face Datasets

Hugging Face (HF) Datasets are currently a key hub for NLP researchers looking for easily usable corpora. The formatting of Pangloss corpora in HF format therefore seemed useful from the perspective of bringing corpora from less-studied languages to the attention of NLP researchers.

Preparing the data for Hugging Face Datasets format Each annotation file is divided into as many files as there are first-level units (sentences for texts, words for vocabulary lists). These files are named after these units, and are placed in folders named after the original resource (see Figure 5b). These data, structured in a tabular manner, contain some relevant metadata, such as the nature of the resource (word list or text) or the speaker. The data are then randomly partitioned. By default, the partitioning is carried out at the first level below the entire resource: hence sentences for texts, and words in the case of word lists. But it is possible to choose the level of the entire file instead. Three sets are thereby created: training, validation and test, thus three CSV files (see figure 7). Finally, all these data (audio files of the sentences and the three CSV files) are uploaded on a server.

Preparing the publicly available dataset A second part of the work consists in formally preparing the new datasets[4]. To achieve this, a precise description of each corpus is required. Particular attention is paid to the encoding of the language concerned[5], a delicate point in the case of languages that are not standardized (or only slightly standardized), which constitute the core business of the Pangloss Collection. A Python script[6] automates the creation of the dataset from the previously uploaded and accessible corpus archives. This script, of varying complexity depending on the dataset, is used to specify where the formatted archives are stored, the data types of each column, and their corresponding names. It is then possible to view the corpus data directly online and to listen to the audio segments (see Figure 8).

The choice made consists in creating one global dataset named Pangloss, of which the various

[4]Available here: https://huggingface.co/datasets/Lacito/pangloss

[5]https://github.com/huggingface/datasets/issues/4881

[6]https://huggingface.co/datasets/Lacito/pangloss/blob/main/pangloss.py

	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_BodyPartsOfA	WORDLIST	Latami, Dashilame	tʂʰɯɪ | ɖɤɪmiɬ-ɬiɬpiJ ɲiJ	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	ŋwɤɪ-qɑ1	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	tʰiʎ, | tʂʰɯɪ | soɪ-bæɬ-hĩɾ | tʂʰɯɪ-dʑoJ, | "tsʰeɬ-jiɬ-dʑoJ | ʑiɬqʰwɤɪ dɑɪ-tsoɪ-ɲiĩ!" | piɬ-zoJ, | gʏJɬiJmiJ-qoɪ-ɳɯJ, | əəə... ɬiJbiJ-tʂʰɯɪ | leɪ-qwæɬ, | wɤʎ | ɑɹʊoɪ | lʏɬqoɪ | tʰiɬ-tʏɬ-hĩɾ-tʂʰɯɪ, | gɤJ-qwæʎ! |	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	tʂʰɯɪ | poɬloɪ-bʏɬ | qʏɬʈsæɬ ɲiĩ	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	gʏɬtsʰiJgʏJ-kʏɬ	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) data for the Hugging Face format

	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_B
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	tʂʰɯɪ | ɖɤɪmiɬ-ɬiɬpiJ ɲiJ	oreilles de renard	狐狸的耳朵	fox's ears	../langues/corpus/Yongnin Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1/Tone_B
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	tʂʰɯɪ | ɖɤɪmiɬ-jiJtʂæJ ɲiJ	taille de renard	狐狸的腰	fox's waist	../langues/corpus/Yongnin Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1/Tone_B
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	tʂʰɯɪ | ɖɤɪmiɬ-njɤJ|ɯJ ɲiJ	yeux de renard	狐狸的眼睛	fox's eyes	../langues/corpus/Yongnin Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1/Tone_B
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	tʂʰɯɪ | ɖɤɪmiɬ-ʉæɬʏɬ ɲiJ	cou de renard	狐狸的脖子	fox's neck	../langues/corpus/Yongnin Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1/Tone_B
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	tʂʰɯɪ | ʐwæɬzoɬ-ɣɯɪ ɲiJ	peau de poulain	马驹子的皮	colt's skin	../langues/corpus/Yongnin Na/corpus HFT Yongning Na – 16k-16/Yongning Na/cocoon-db3cf0e1-30bb-3225-b012-019252bb4f4d_C1/Tone_B
9404	cocoon-db3cf0e1-	Yongning Na – 16k-	cocoon-	cocoon-	Tone_BodyPartsOfAnim	WORDLIST	Latami,	280.25	282.25	tʂʰɯɪ | ʐwæɬzoɬ-bʏɬ ɲiĩ	intestin de	马驹子的场子	colt's	../langues/corpus/Yongnin

(b) data in Hugging Face format, after random partitioning

Figure 7: OutilsPangloss: CSV files

languages (Japhug[7], and Yongning Na[8]) constitute sub-sets.

It is then possible to load the desired data set and get experiments started in just two lines of Python code.[9]

[7] https://huggingface.co/datasets/Lacito/pangloss/viewer/japh1234/train

[8] https://huggingface.co/datasets/Lacito/pangloss/viewer/yong1288/train

[9] `import datasets` then `datasets.load_dataset("pangloss", "japh1234")` for Japhug, and `datasets.load_dataset("pangloss", "yong1288")` for Yongning Na.

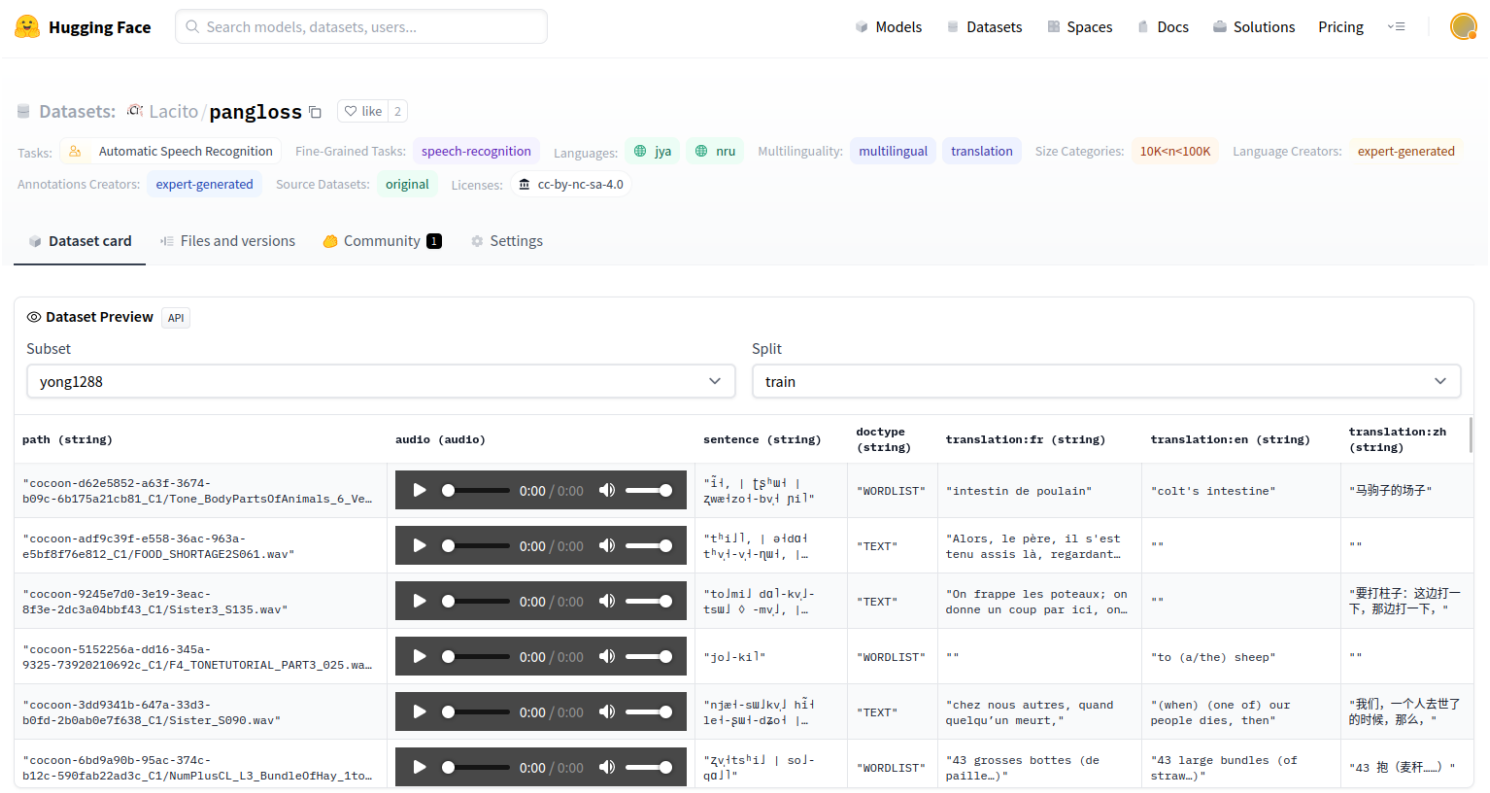


Figure 8: Page of the Pangloss corpus on Hugging Face Datasets

5 Perspectives: what is at stake in providing a full description of datasets without creating "hard copies"

The ongoing transition to Open Science carries a fundamental requirement for reproducibility of experiments, and the field of speech sciences and Automatic Language Processing is no exception (Garellek et al., 2020). Our hope is that practices will gradually evolve towards a description of datasets via metadata pointing to one master file hosted in an archive that guarantees both long-term conservation and 24/7 online availability. Describing the datasets in this way only takes a few kilobytes (Kb), whereas hosting each dataset as a "hard copy" would result in multiple and highly redundant deposits (in Zenodo or elsewhere) each of which amounts to gigabytes (GB).

Acknowledgments

Many thanks to the language consultants and friends for Japhug (in particular Tshendzin) and Na (in particular Mrs. Latami Dashilame and her son Latami Dashi). The present work is a contribution to the project "Computational language documentation by 2025" (ANR-19-CE38-0015-04) and to the Labex "Empirical foundations of linguistics" (ANR-10-LABX-0083). We thank the Institute for Linguistic Heritage and Diversity (ILARA) at *École pratique des hautes études*, the University of Queensland and the Australian Research Council Centre of Excellence for the Dynamics of Language for financial support to software development for language documentation.

References

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., NGUYÊN, 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., NGUYÊN, 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

KIK, A., ADAMEC, M., AIKHENVALD, A. Y., BAJZEKOVA, J., BARO, N., BOWERN, C., COLWELL, R. K., DROZD, P., DUDA, P., IBALIM, S., JORGE, L. R., MOGINA, J., RULI, B., SAM, K., SARVASY, H., SAULEI, S., WEIBLEN, G. D., ZRZAVY, J., & NOVOTNY, V. (2021). Language and ethnobiological skills decline precipitously in papua new guinea, the world's most linguistically diverse nation. *Proceedings of the National Academy of Sciences*, *118*(22). https://doi.org/10.1073/pnas.2100096118

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