PRESS RELEASE - PARIS - 14 JANUARY 2021

Endangered linguistic heritage: a new website for the Pangloss Collection

Like certain animal and plant species, some of the world's languages are in danger of extinction. Fortunately, the Pangloss Collection, an open archive started in 1995 by the *Langues et civilisations à tradition orale* laboratory (CNRS/Université Sorbonne Nouvelle/Inalco), makes available recordings of endangered languages in order to preserve this linguistic heritage and make it open-access. Languages without a written tradition (the vast majority) could otherwise disappear completely when their last speakers pass away. Other relatively undocumented languages are also included in the collection. Thanks to the support of the CNRS, the Pangloss Collection is now being revamped with a new website, also accessible to the general public.

To date, the Pangloss open archive contains more than 3600 audio and video recordings in 170 languages from across all continents. For instance, it includes stories and songs in Xârâgurè (New Caledonia), conversations and tales in Kakabe (Guinea) and cooking recipes in Koyi rai (Nepal) and Na-našu (Italy) – a total of 780 hours of listening.

These data are the result of more than twenty years' work by linguists and ethnologists who, in their own field of study, are working to collect and preserve the world's linguistic heritage. Some of the documents come from the digitisation of old magnetic tapes[1]. Nearly half of the recordings are transcribed and annotated, some with contextual elements or translations into other languages. The site is open to contributions from both academic and non-academic experts, who are encouraged to improve the corpus by contributing to transcriptions and translations.

In order to be more accessible to the general public, who can freely listen to and download these precious documents and hereby get a sense for the world's linguistic diversity, the redesigned pangloss.cnrs.fr website can now be consulted via two levels of access. As the content is largely under a *Creative Commons* licence, it is available for use in museographic projects or audio creations.

Beyond its heritage aspect, this collection is also part of an open science approach to facilitate the conservation, referencing[2] and availability of primary data for researchers. Its purpose is to limit the loss of scientific data (a "second death" for extinct languages) whilst also encouraging collaboration with other disciplines: computer scientists interested in automatic language processing can access the files they need and take part in the co-development of tools (e.g. for automatic transcription). The site is fully bilingual (French–English) and also includes partial translations in other languages, including Chinese for records in certain Asian languages.

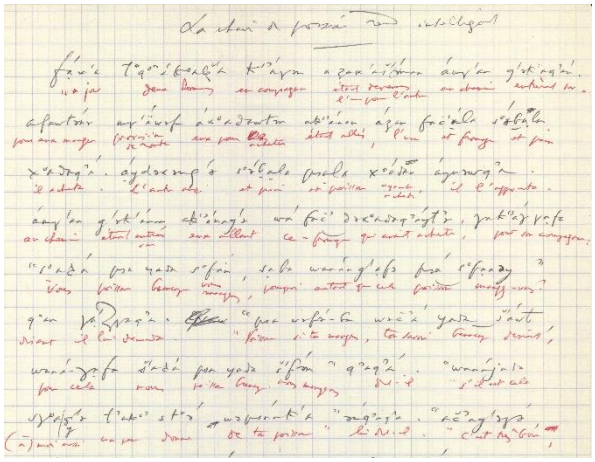
*In addition to contributions from various laboratories associated with the CNRS[3], the Pangloss Collection is supported by the recently created Institute for Linguistic Heritage and Diversity at the EPHE-PSL, and data are stored in the archive of the large research infrastructure (*Très grande infrastructure de recherche *–TGIR) Huma-Num. The Pangloss Collection is a member of the international Digital Endangered Languages and Musics Archives Network (DELAMAN). It is hosted by the Cocoon platform,* Collection de corpus oraux numériques*, which is one of the participating archives of the Open Language Archive Community (OLAC).*

Notes

[1] Such as those of the last speaker of the Ubykh language, by Georges Dumézil, in 1968.
[2] Each resource has DOI (Digital Object Identifier) and ARK (Archival Resource Key) identifiers.
[3] In particular (non-exhaustive list): *Langues et civilisations à tradition orale* (Lacito, CNRS/Université Sorbonne Nouvelle/Inalco); *Centre de recherches linguistiques sur l'Asie orientale* (CRLAO, CNRS/Inalco/EHESS); *Langage, langues et cultures d'Afrique noire* (Llacan, CNRS/Inalco); *Structure et dynamique des langues* (Sedyl, CNRS/Inalco/IRD).

Some examples from the website:

- 'Eating Fish Makes You Clever', a story in Ubykh (a Caucasian language formerly spoken in Turkey and Georgia, which has some 80 consonants) told by Tevfik Esenç, its last speaker, and recorded by the linguist and anthropologist Georges Dumézil in 1968 (his handwritten notes can also be consulted) doi.org/10.24397/pangloss-0004320
- Audio and video corpus in Kakabe, a language of Guinea (that used to be a language of slaves or servants), recorded and deposited by linguist Alexandra Vydrina: pangloss.cnrs.fr/corpus/Kakabe?lang=en



Georges Dumézil's handwritten transcription of an Ubykh story, as told by Tevfik Esenç.
© Georges Dumézil

A group of Kouroupampa (Guinea) women cooking (photo from the Kakabe language corpus). © Alexandra Vydrina

Contacts

CNRS Researcher I Alexis Michaud I alexis.michaud@cnrs.fr
CNRS Engineer I Séverine Guillaume I severine.guillaume@cnrs.fr
CNRS Press Officer I Véronique Etienne I **T +33 1 44 96 51 37** I veronique.etienne@cnrs.fr