HAL open science

”La Collection Pangloss : une archive ouverte de langues peu documentées, conçue pour favoriser une utilisation en Traitement Automatique des Langues”

Séverine Guillaume, Balthazar Do Nascimento, Alexis Michaud

▶ **To cite this version:**

Séverine Guillaume, Balthazar Do Nascimento, Alexis Michaud. ”La Collection Pangloss : une archive ouverte de langues peu documentées, conçue pour favoriser une utilisation en Traitement Automatique des Langues”. Conférence internationale sur les technologies linguistiques pour tous (LT4AII) : Favoriser la diversité linguistique et le multilinguisme dans le monde (UNESCO), Dec 2019, Paris, France. , 2019. hal-02415045

HAL Id: hal-02415045

https://hal.science/hal-02415045v1

Submitted on 16 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

La Collection Pangloss

Séverine Guillaume, Balthazar Do Nascimento, Alexis Michaud

Collection Pangloss :

Langues principalement à tradition orale
170 langues, **745 heures d'enregistrements**
(720h d'audio et 25h de vidéo),
1700 documents (230h) **transcrits, traduits, glosés**

Documents recueillis par des linguistes de terrain «généralistes» dans le monde entier

- **Documentation fondamentale** pour la sauvegarde, diffusion et valorisation de langues minoritaires, en danger

- **Pérennisation** de documents uniques, fruits d'explorations non reproductibles (archivage pérenne au CINES via les services de la TGIR Huma-Num)

- Accueille des données de **langues de toutes les aires géographiques et culturelles** (environ 20 langues en 2001, 70 langues en 2011, 170 à ce jour)

Extrait XML des annotations de l'enregistrement audio

```
<TEXT xml:lang="lhu" id="crdo-LHU_LAHU_JM_20">
  <HEADER>
    <TITLE xml:lang="en">How We Came from Burma</TITLE>
    <SOUNDFILE href="Thailand/Lahu/LAHU_JM_20.wav"/>
  </HEADER>
   (...)
   <S id="LAHU_JM_20s3">
    <AUDIO start="15.9382" end="20.0724"/>
    <FORM kindOf="phono">yâ kà? ɔ-mō=ló mâ ve... tô ve...</FORM>
    <TRANSL xml:lang="en">And there was a big flock of kids along too--er, walking along--.</TRANSL>
    <W>
      <FORM kindOf="phono">yâ</FORM>
      <TRANSL xml:lang="en">children</TRANSL>
      <M>
        <FORM kindOf="phono">yâ</FORM>
        <TRANSL xml:lang="en">children</TRANSL>
      </M>
    </W>
    <W>
      <FORM kindOf="phono">kà?</FORM>
      <TRANSL xml:lang="en">PRT(also)</TRANSL>
      <M>
        <FORM kindOf="phono">kà?</FORM>
        <TRANSL xml:lang="en">PRT(also)</TRANSL>
      </M>
    </W>
    <W>
      <FORM kindOf="phono">ɔ-mō=ló</FORM>
      <TRANSL xml:lang="en">big.group</TRANSL>
      <M>
        <FORM kindOf="phono">ɔ</FORM>
        <TRANSL xml:lang="en">PREF</TRANSL>
      </M>
      <M>
        <FORM kindOf="phono">mō</FORM>
        <TRANSL xml:lang="en">group</TRANSL>
      </M>
      <M>
        <FORM kindOf="phono">ló</FORM>
        <TRANSL xml:lang="en">big</TRANSL>
      </M>
    </W>
    (...)
  </S>
</TEXT>
```

Interface de consultation en ligne

Technologies de la parole pour langues peu dotées : un domaine de recherche très actif

- Enrichir la base de données au moyen d'outils numériques
  - Alignement temporel texte-son par alignement forcé
  - Etiqueteurs morphosyntaxiques pour génération automatique de gloses
- Défi scientifique du scénario « ressources limitées » pour la recherche en informatique
  - Transcription automatique de la parole
    Initiatives récentes : `wav2letter++, persephone, elpis...`
  - Synthèse de la parole
  - Traduction automatique
- Potentiel pour l'enseignement : outils ludiques grand public

https://pangloss.cnrs.fr/