VERA - Vidéo, Ethnotextes et Ressources Associées

Pierre-Aurélien Georges

VERA

Vidéos, Ethnotextes et Ressources Associées

http://thesaurus.unice.fr/vera

Pierre-Aurélien Georges
pageorge@unice.fr

THESOC — BASES CORPUS LANGAGE — UMR 7320 CNRS / UNS / UCA

Contexte

Des enregistrements inédits…

2006 – 2017 :

+ de 100 enquêtes de terrain

Dialectes occitans et liguriens
de 75 localités (France & Italie)

5 questionnaires de phrases
totalisant 1200 questions

Bilan de ces 9 campagnes d'enquêtes :

+ de 120 locuteurs interviewés
58 ethnotextes récoltés
60 h de vidéos inédites à traiter
(et autant d'enregistrements audio)
env. 400 000 occurr. à lemmatiser

Traitements

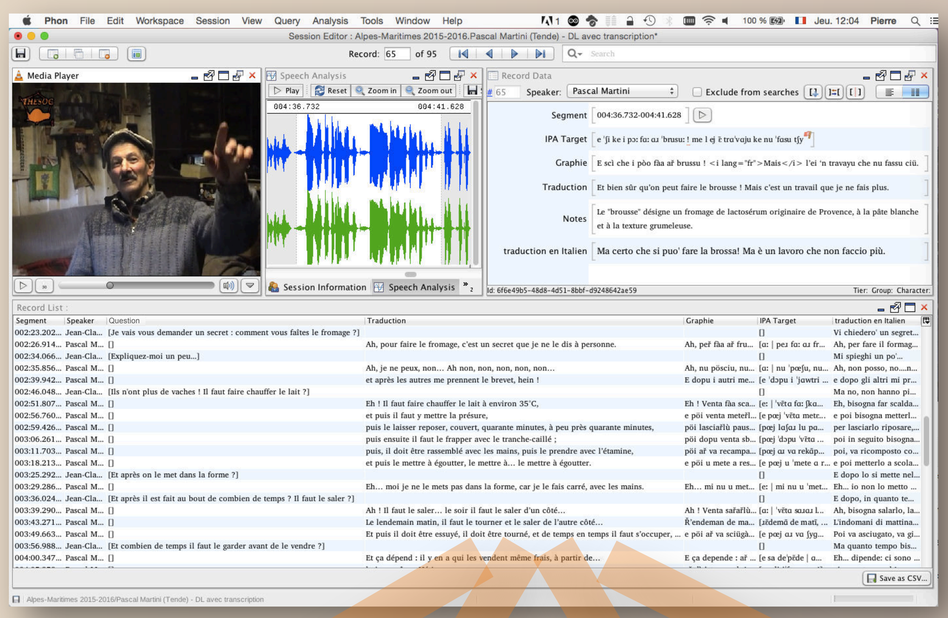
…accompagnés de nouvelles ressources…

1) Montage Vidéo

Normalisation du volume sonore,
Synchronisation audio – vidéo,
Choix des séquences vidéos,
Titres et logos.
(Adobe Premiere)

2) Segmentation

Repérage des tours de parole,
Minutage phrase par phrase,
Numérotation des réponses
aux questionnaires de phrases
(PHON.ca – XML PhonBank)



3) Transcriptions

en graphie et en Alphabet Phonétique International
- ou - dans 2 graphies différentes, suivant les enquêtes.

4) Traduction

alignée phrase par phrase, en français et parfois italien,
avec commentaires et explication des termes opaques.

5) Annotations

Lemmatisation et étiquettes P.O.S.
(base MMS du Thesoc)

6) Encodage XML-TEI et géolocalisation

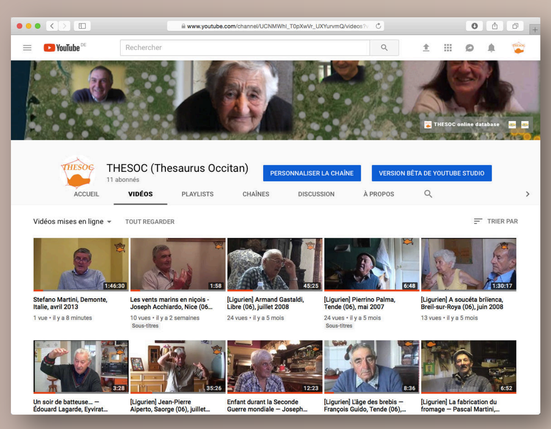
(métadonnées Dublin-Core et OLAC)

Livrables

…consultables selon 4 modalités différentes…

Chaîne vidéo YouTube™

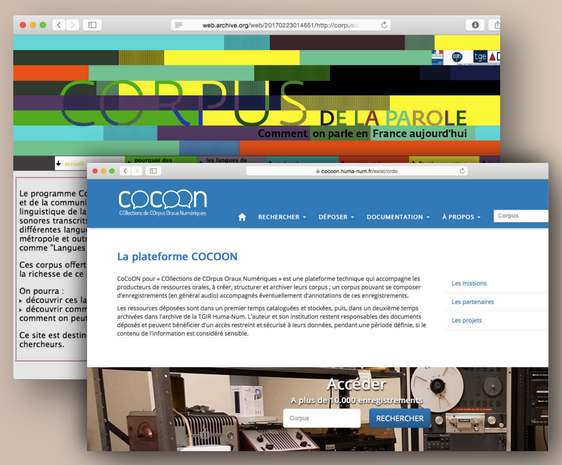
permettant au grand public de
découvrir tous nos discours libres



https://bit.ly/thesOC

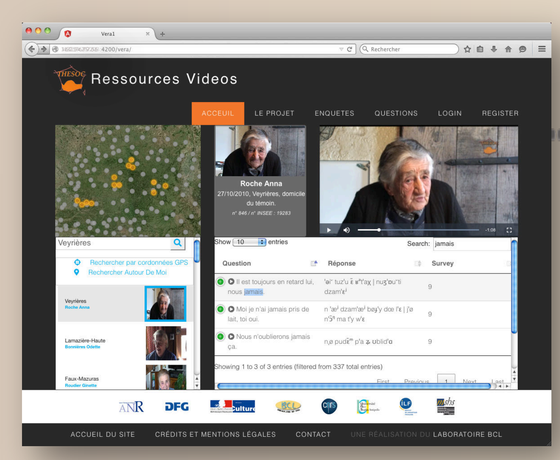
Archivage pérenne

et versement au ***Corpus de la Parole***



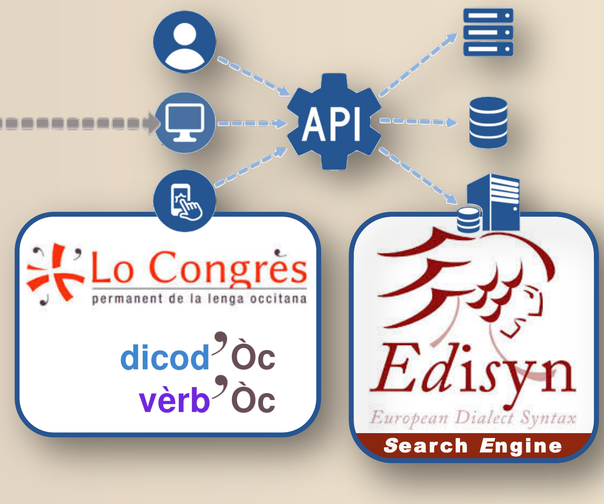
Plateforme web dédiée

avec cartographie interactive
et fonctions d'exploration du corpus



Interface de requête **API REST**

(interopérabilité, notamment avec les
moteurs de recherche spécialisés)



Exploitations

…avec de multiples applications possibles :

Industries de la Langue : corpus d'entraînement en traduction automatique, Text2Speech, reconn. vocale, …
Enseignement des langues régionales : support pédagogique pour les enseignants et étudiants
Linguistique : lexique, phonologie, microvariation syntaxique, toponymie, …
Communication non verbale : analyse de la gestuelle et expressions du visage
Histoire, sociologie, ethnologie (ethnotextes aux contenus très variés)

Ressources linguistiques sous licence LGPL-LR – Vidéos sous licence CC-BY-NC-ND

CEPD'OC Centre d'Etudes de la Parole D'OC cepd-oc.fr
Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE Ministère Culture
Huma-Num
CORpus Langues Interactions
cnrs
INVESTISSEMENTS D'AVENIR INITIATIVE D'EXCELLENCE
UCA J.E.D.I. UNIVERSITÉ CÔTE D'AZUR
Ce travail a bénéficié d'une aide du gouvernement français, gérée par l'Agence Nationale de la Recherche au titre du projet Investissements d'Avenir UCAJEDI portant la référence n° ANR-15-IDEX-01