

# Cendres de Ressources pour les Données Orales (CRDO)

*Rapport intermédiaire, Juin 2006*

## 1. Organisation

Le CRDO a constitué dès sa création un comité d'organisation formé de représentants de ses deux composantes (5 personnes au total, dont un DR et 5 ingénieurs). Pour des raisons d'efficacité et compte tenu des délais impartis, nous avons décidé de distinguer deux sous-projets. Le premier, conduit par le LPL à Aix, porte plus particulièrement sur le format et le traitement des données orales ainsi que les outils nécessaires à leur création. L'autre projet, sous la responsabilité du LACITO à Paris, porte plus particulièrement sur les ressources documentaires. Chacun de ces sous-projets fonctionne de façon autonome, la coordination étant assurée par le comité d'organisation. Les objectifs respectifs étant distincts, chaque projet s'est doté d'un comité de pilotage qui aura pour tâche de conduire à très court terme un ensemble de recommandations et proposer des échantillons de corpus. Le résultat visé est la mise au point d'un prototype de site pour la fin de cette année. Ce prototype sera l'addition des contributions des deux projets.

## 2. Activités du pôle Aixois

### 2.1. Mise en place du Comité de Pilotage

Le succès du CRDO passe par l'implication de la communauté à tous les niveaux de son organigramme. Nous avons donc créé un comité de pilotage formé de quelques uns des principaux acteurs dans le domaine de la production de données orales. Ce sont les laboratoires BCL, CORAL, ERSS, GIPSA (ex-ICP), ICAR, LIA, LIMSI, LORIA, LPL, LPP, MODYCO auxquels se joint l'ATILF au titre du CRNTL. Nous assurons ainsi une couverture large des travaux dans ce domaine. Chacun de ces partenaires dispose d'ores et déjà d'une compétence soit en tant qu créateur de données et d'outils, soit en tant qu'utilisateur.

### 2.2. Travaux effectués

La première étape a consisté à collecter un ensemble de données orales brutes ou enrichies par des annotations. Nous avons pour cela commencé à travailler sur un ensemble varié de données existantes, en identifiant le type d'information pouvant leur être associé, en proposant un encodage XML respectant les standards internationaux. Ce travail est le fruit d'une collaboration de plusieurs des membres du comité de pilotage. Nous avons en particulier commencé à tester cette approche sur un ensemble de corpus multimodaux, les données orales étant en effet souvent associées à d'autres modalités. Cette réflexion a été initiée dans le cadre d'un projet ILF. Elle s'est élargie à d'autres partenaires pour déboucher sur la proposition d'un projet ANR dont le LPL est porteur.

L'objectif à court terme est de traiter les points suivants :

- Etat de l'art des normes et standards pour le traitement des données linguistiques orales
- Développement d'une recommandation d'annotation pour ces données
- Application à la normalisation de ressources existantes
- Déploiement d'outils de requête sur les données et les métadonnées
- Déploiement du site web pour la mise à disposition avancée de ces données

### 2.3. Prochaines étapes

La prochaine réunion du comité de pilotage (12 septembre à Lyon) sera l'occasion d'une part de faire le point du projet et d'autre part, pour chacun des partenaires, de verser au site un ensemble de données orales dont il est détenteur. Nous disposerons donc à terme d'un échantillon très varié de corpus destinés originellement à des recherches très différentes allant de la phonétique à l'analyse des interactions en passant par l'étude de la parole pathologique. Par ailleurs, plusieurs équipes membres du comité de pilotage

sont également développeurs d'outils d'aide à la création et l'enrichissement de ressources orales. Ceux-ci auront également pour vocation à être diffusés.

Dès la fin de l'année, tous ces échantillons, ainsi qu'une méthode d'investigation mais également des outils d'aide à la création de ressources seront disponibles via cette branche du CRDO.

### **3. Activités du pôle Parisien**

#### ***3.1. Contact Européen***

Une mission a été organisée en mars 2006 pour visiter l'équipe technique de Peter Wittenburg au PMI de Nimègue. Cette visite a débouché sur la volonté de part et d'autre de trouver des terrains de coopération. Un premier terrain pourrait être pour nous de délivrer les métadonnées dans le schéma IMDI, pour eux de moissonner les archives avec le protocole OAI. Nous avons aussi signé avec eux la charte sur les « live archives »<sup>1</sup>

#### ***3.2. Contacts Français***

Participation à l'école thématique CNRS « Linguistique de Corpus Oraux » de Nantes (19 au 24 juin 2006). De nombreux contacts ont pu être pris à cette occasion avec l'ensemble de la communauté. Ces contacts devraient déboucher rapidement sur plusieurs projets de dépôt et de valorisation de corpus.

#### ***3.3. Dossiers ANR***

Quatre dossiers ANR (*Corpus et outils de la recherche en sciences humaines et sociales*) ont été fait mentionnant le CRDO soit au titre de partenaire (sur des ressources en français avec un aspect standardisation des codages) soit au titre de prestataire (un sur les outils, un sur des ressources orales de Nouvelles Calédonie, un dernier sur des ressources orales des langues de la famille Afroasiatique)

#### ***3.4. Productions et mises à disposition***

Un site web a été créé (mise en activité février 2006) pour présenter la composante parisienne du crdo et servir de portail sur les ressources (<http://crdo.vjf.cnrs.fr>). Ce site présente l'organisation, les missions et les projets en cours de cette composante. Il offre aussi une interface d'accès aux ressources. Par ailleurs, un nom a été demandé à Renater ([www.crdo.fr](http://www.crdo.fr)) afin de fédérer les 2 composantes du centre. Nous attendons toujours leur réponse.

Un réservoir de ressources a été créé (en février 2006) basé entièrement sur les technologies W3C. Un programme implémentant le protocole OAI-PMH a été développé afin de disséminer les métadonnées. Il a été enregistré auprès des organisations OAI et OLAC comme « data provider ». Un moteur de recherche dans le catalogue des ressources a été développé, ainsi qu'une interface de consultation des ressources. Ce réservoir contient à ce jour 75 heures d'enregistrements dont 20 heures en libre accès. Afin de couvrir l'ensemble des besoins exprimés par les chercheurs, nous avons distingué 3 niveaux de protection : 1) données et métadonnées en libre accès, 2) métadonnées publiques mais données accessibles par authentification, 3) données et métadonnées accessibles par authentification.

Les archives du Lacito ont été versées en totalité dans ce réservoir<sup>2</sup>. De nouvelles ressources ont été déposées par l'Université d'Orléans<sup>3</sup>, l'Université de Nantes<sup>4</sup>, le CELIA<sup>5</sup>, et le LACITO<sup>6</sup>. De nombreuses heures d'enregistrement sont aussi en cours de dépôt (juin) notamment un corpus en judéo-hispanique (LMS) et un autre corpus en maoré (LACITO).

L'Université d'Orléans a aussi déposé un extrait de l'enquête ESLO (environ 80 enregistrements correspondants aux réponses à une des questions de l'enquête). Nous avons développé pour cette enquête une maquette de consultation mixant dans les requêtes des critères socio-linguistiques et des critères portant sur la transcription.

---

1 [http://www.mpi.nl/DAM-LR/flyers/DLRA\\_Flyer\\_2006-04-23.pdf](http://www.mpi.nl/DAM-LR/flyers/DLRA_Flyer_2006-04-23.pdf)

2 Une trentaine de langues

3 Langues: angolais, saotomense et français

4 Du kabyle

5 Du ndyuka

6 De nombreuses langues de Nouvelle-Calédonie

### *3.5. Projets en cours*

Les corpus qui vont être déposés avant l'automne :

- L'ensemble des corpus numérisés dans le cadre du contrat entre la DGLFLF et la fédération TUL pour un portail sur le français et les langues de France pour le ministère de la Culture.
- L'enquête Socio-Linguistique d'Orléans (ESLO) soit 350 heures d'enregistrement de Français des années 70 (Université d'Orléans)
- Un corpus de créole et de français d'une centaine d'heures de (Université de l'île de la Réunion).
- Corpus kabyle d'Algérie (Université de Nantes)
- Créoles et langues de Guyane (Université d'Orléans)
- Un Corpus de Français d'Abijan (Université de Besançon)
- Plusieurs corpus de langues d'Afrique (LLACAN)
- Corpus tibétain d'une centaine de documents (contrat PICS avec l'Université de Virginie).
- des contacts avancés sont en cours avec le projet PFC (ERSS, MoDyco, Universités de Nanterre, de Toulouse d'Oslo et de Tromsø)

Un comité de pilotage réunira les représentants de tous les laboratoires de la composante parisienne du CRDO + un représentant de la composante LPL-Aix + 2 experts extérieurs (ICAR et CORAL).

Notre groupe s'active fortement dans la perspective de mise à disposition de corpus existants. Un important travail a déjà été effectué pour la mise en place d'une architecture de stockage et diffusion de ces corpus qui respectent au plus près les standards internationaux. Cette architecture est déjà opérationnelle et nous permettra d'accueillir facilement les nouveaux corpus de toute la communauté.