Presentation on CLARIN in France (CLARIN-FR)

Patrice Bellot (Université Aix-Marseille)
Nicolas Larrousse (Huma-Num)

patrice.bellot@univ-amu.fr / nicolas.larrousse@huma-num.fr
CLARIN-FR Roadmap

- **Observer** in CLARIN with the endorsement of the FR Minist. of Research

- French situation in 2016:
  - A *very large facility* dedicated to Humanities and Social Sciences that leads the participation of France in CLARIN and in DARIAH
    - Huma-Num
  - and a network infrastructure for Language data and tools:
    - Ortolang

- The first aim for France in CLARIN:
  - to gather the French structures already involved (two C centres in 2016) + Ortolang + some labs
  - to propose a road map before becoming a full CLARIN member (economic model, existing scientific communities)

## Speech & Language Data Repository

<table>
<thead>
<tr>
<th>Website</th>
<th><a href="http://sldr.org/">http://sldr.org/</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Consortium</td>
<td>none</td>
</tr>
<tr>
<td>Type(s)</td>
<td>C</td>
</tr>
<tr>
<td>Description</td>
<td>SLDR is a repository for oral and linguistic resources aimed at their long-term preservation and sharing</td>
</tr>
<tr>
<td>DSA</td>
<td>none</td>
</tr>
<tr>
<td>PID status</td>
<td>Handle</td>
</tr>
<tr>
<td>Repository system</td>
<td>Self-developed</td>
</tr>
<tr>
<td>Strict versioning?</td>
<td>✓</td>
</tr>
</tbody>
</table>

### Organisational information

<table>
<thead>
<tr>
<th>Organisation name</th>
<th>Laboratoire Parole et Langage (LPL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Institution</td>
<td>Centre National de la Recherche Scientifique (CNRS) and Aix-Marseille University</td>
</tr>
<tr>
<td>Working unit</td>
<td>UMR 7309</td>
</tr>
<tr>
<td>Shorthand</td>
<td>SLD R</td>
</tr>
<tr>
<td>Postal address</td>
<td>5 avenue Pasteur 13100 Aix-en-Provence France</td>
</tr>
<tr>
<td>Expertise</td>
<td>oral and linguistic resources</td>
</tr>
</tbody>
</table>
Deposit and sharing of oral/multimodal linguistic data

As of 01/12/2015, registration of new users and deposit of data on SLDR website will be suspended to allow the public opening of Ortolang platform.

Now engaged with the CNRTL and Nanterre Orléans Centre in building ORTOLANG, SLDR continues its work of gathering and sharing language data. All services currently offered will remain part of the new platform.

Thus, SLDR/ORTOLANG allows you to browse data already collected in the area of speech/multimodal linguistics. Resources are grouped into four main types of items: primary data, secondary data, tools and collections. Downloading is possible on the basis of various provisions of archival law and intellectual property rights.

After authenticating on the site, you can also deposit your data and create their descriptive notice via a dedicated interface. This will notably allow you to specify access to data based on its status, nature, or professional categories of users.

Descriptive records created are consistent with the core metadata standards: they are eligible for harvesting by international directories and search engines. Meanwhile, items and their contents are assigned persistent identifiers to facilitate their retrieval regardless of physical location.

These steps are intended to finally prepare items for their long-term preservation by CINES, an institutional archive site.

Do not hesitate to contact us for additional information, or read our guidelines for the sharing and archiving of linguistic resources.

The latest deposits (169) >> more

Primary data (corpus) ortolang-000939
SITAF (tandem anglais/français) (Céline HORGUES)
Individual contribution

Video-recorded corpus of tandem interactions between French-speaking students and English-speaking students at the University Sorbonne Nouvelle-Paris 3. Each of the 21 tandem pairs performs the collaborative speaking tasks (story-telling, debating, reading) on 2 occasions (2 recording sessions) 3 months apart. The corpus also comprises LI-L1 control interactions for the participants.

Metadata is [...]

(applied linguistics, language acquisition, phonetics)
French; English

Primary data (corpus) bhr-000940
Projet Noms Verbos (MEG) (ALASDIFrancois-Xavier, LPC, BADIER Jean-Michel, INS, STRIJKERS Kristof, LPC, CHANOINE

(hdl:11411/orlang-000939
2015-09-30
Version 1
source data
(© Misc publications)
Collections de corpus oraux numeriques

<table>
<thead>
<tr>
<th>Website</th>
<th><a href="http://cocoon.huma-num.fr">http://cocoon.huma-num.fr</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Consortium</td>
<td>none</td>
</tr>
<tr>
<td>Type(s)</td>
<td>C</td>
</tr>
<tr>
<td>Description</td>
<td>Speech recordings repository</td>
</tr>
<tr>
<td>DSA</td>
<td>none</td>
</tr>
<tr>
<td>PID status</td>
<td>ARK and PURL</td>
</tr>
<tr>
<td>Repository system</td>
<td>self-developed</td>
</tr>
<tr>
<td>Strict versioning?</td>
<td>✗</td>
</tr>
</tbody>
</table>

Organisational information

<table>
<thead>
<tr>
<th>Organisation name</th>
<th>CNRS/COCOON</th>
</tr>
</thead>
<tbody>
<tr>
<td>Institution</td>
<td>Centre National de la Recherche Scientifique</td>
</tr>
<tr>
<td>Working unit</td>
<td>Laboration ligérien de linguistique (LLL) &amp; Langues et civilisations à tradition orale (LACITO)</td>
</tr>
<tr>
<td>Shorthand</td>
<td>COCOON</td>
</tr>
<tr>
<td>Postal address</td>
<td>Bibliothèque nationale de France, Quai François Mauriac, 75013 Paris 75013 Paris France</td>
</tr>
</tbody>
</table>
COollections de COorpus Oraux Numériques

COCOOn est une plateforme technique qui accompagne les producteurs de ressources orales, à créer, structurer et archiver leurs corpus ; un corpus pouvant se composer d’enregistrements (en général audio) accompagnés éventuellement d’annotations de ces enregistrements.

Les ressources déposées sont dans un premier temps cataloguées et stockées, puis, dans un deuxième temps archivées dans l’archive de la TGIR Huma-Num. L’auteur et son institution restent responsables des documents déposés et peuvent bénéficier d’un accès restreint et sécurisé à leurs données, pendant une période définie, si le contenu de l’information est considéré sensible.

La plateforme COCOON est gérée conjointement par deux unités mixtes de recherche : le Laboratoire de Langues et civilisations à tradition orale (LACITO - UMR7107 - Université Paris3 / INALCO / CNRS) et le Laboratoire Ligérien de Linguistique (LLL - UMR7270 - Universités d’Orléans et de Tours, BnF, CNRS).

Statistiques

Actuellement sont consultables dans l’archive : 9019 enregistrements pour un total de 153 langues différentes.
During the conference

• You can meet some of the French labs / structures that may (wish to) become CLARIN centres:

  • LIMSI (CNRS - Paris Orsay)
  • LINA (Université de Nantes)
  • LPL, LIF-LSIS (Université Aix-Marseille)
  • Ortolang
  • ELDA
**Termsuite**

Termsuite is a Java UIMA-based toolbox for terminology extraction and multilingual term alignment. It extracts monolingual terminologies and generates bilingual dictionaries from these terminologies by the means of distributional and compositional methods.

The languages covered are: English, French, German, Spanish, and Russian.

[More information here](https://www.lina.univ-nantes.fr)

**Apopsis**

Un détecteur d'opinions qui explore les tweets sur le sujet qui vous intéresse ! Une personne, un produit, un sujet d'actualité. Entrez le sujet et visualisez en temps réel les opinions qui sont émises positives, négatives ou neutres.

[ Jouer avec un démonstrateur ici](https://www.lina.univ-nantes.fr)

**Apache OpenNLP models for processing French**

The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text. It supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference resolution.

[More information here](https://www.lina.univ-nantes.fr)

**Code source and Apache UIMA components**

Various contributions to the NLP and Apache UIMA (Unstructured Information Management Architecture) communities to facilitate the development of NLP components and pipelines, to connect with various data formats, to solve interoperability issues (within UIMA workflow or by integrating third-party tools) and also to perform some NLP analysis tasks.

Here two source code repositories dev-star et jules-star ; both should be merged soon. Some components are available under [Java Apache Maven dependencies](https://www.lina.univ-nantes.fr).
Modeling Language

- Speech Recognition
- Textual Documents
- Sign Language
- Speaker Characterization
- Detection of Affects, Sentiments, Opinions
- Language Variation
- Paraphrasing
- Machine Translation
- Machine Learning
- Information Extraction
- Question-Answering
- Human-Machine Dialogue

* ILES: Written and Signed Language Information Processing
* TLP: Spoken Language Processing

https://www.limsi.fr/

Gilles Adda
The Evaluations and Language resources Distribution Agency

ELDA was created parallel to ELRA, European Language Resources Association, in February 1995.

ELDA is the association’s operational body, and is in charge of the development and the execution of ELRA’s missions and tasks as defined by the Board of the association.

ELDA is incorporated as a company in order to handle all the commercial and business-oriented tasks of the association.

The CEO of the agency, Khalid Choukri, also acts as ELRA Secretary General and currently works with 12 full-time and permanent members. Some temporary staff may join from time to time to help on some specific projects.
Infrastructure

VLRI (Very Large Research Infrastructure)?

European level … and further

ERICs (European Research Infrastructure Consortium)
Huma-Num’s Ecosystem

SUPPORTING RESEARCH COMMUNITIES

CONSORTIA
Digitalisation
Good practices
Expertises

TOOLS AND SERVICES
Store. Process. Disseminate
Share&Display. Tag&PUSH. Archive

RESEARCH PROJECTS

INTERNATIONAL NETWORKS
Sharing. Expertise
Huma-Num’s Consortia

CAHIER
(Corpus d’Auteurs pour les Humanités : Informatisation, Édition, Recherche)

ARCHIPOLIS
(Archives des sciences sociales du politique)

CONSORTIUM MUSICA

IMAGEO
(Cartes et photographies pour les géographes)

CONSORTIUM ARCHIVES DES ETHNOLOGUES

CORLI
( Corps, langues, interactions)

3D-SHS

MASA
(Mémoires des archéologues et des sites archéologiques)

COSME
(Consortium sources médiévales)

ARCMC
(Archives des mondes contemporains)
Huma-Num’s Core Services

SERVICES FOR DIGITAL RESEARCH DATA

- **STORE**
  - Preserve. Organize

- **ARCHIVE**
  - Long term preservation

- **PROCESS**
  - Tools. Softwares

- **TAG & PUSH**
  - Semantic enrichment
  - Unified access

- **DISSEMINATE**
  - Virtual machines
  - Web diffusion

- **SHARE & DISPLAY**
  - Document
  - Show through Nakalona

Partnerships with CC-IN2P3 and CINES
In a few words…

- Observer for the next 2 or 3 years….
- … even if we are still waiting for the signature
- Candidatures for new CLARIN Centres to come

patrice.bellot@univ-amu.fr / nicolas.larrousse@huma-num.fr