# CORLI: a linguistic consortium for corpus, language and interaction

**Christophe Parisse**
Modyco
Inserm, University of Nanterre, France
cparisse@u-paris10.fr

**Céline Poudat**
BCL
Université Côte d'Azur, France
poudat@unice.fr

**Ciara Wigham**
Laboratoire de Recherche sur le Langage
Université Clermont Auvergne, France
ciara.wigham@uca.fr

**Michel Jacobson**
LLL,
Université d'Orléans et Tours, France
michel.jacobson@gmail.com

**Loïc Liégeois**
CLILLAC-ARP (EA 3967) et LLF
Université Paris Diderot, France
loic.liegeois@univ-paris-diderot.fr

**Abstract**

CORLI is a consortium of Huma-Num, an organization that helps to organize and provide services for digital humanities in France. CORLI is the consortium for linguistics and includes all aspects of linguistic research and development.
As France just joined Clarin as an observer, the objective of our paper is to introduce the consortium CORLI to Clarin; CORLI will act as an interface between Clarin and the scientific community of linguists.
The goal of CORLI is to help linguists create, use, and disseminate linguistic corpora and digital tools. CORLI has always maintained a policy of providing funding and technological help to finalize and publish corpora issued from a wide range of institutional or personal research projects. CORLI is also involved in recommending and the circulation of guidelines related to research and technical practices, especially about linguistic corpora. Finally, CORLI organizes workgroups whose goal is to create and moderate networks that target tools and practices in linguistics. These workgroups are organised thematically around topics including metadata, formats, tools and practices for corpus exploration, archiving systems, multimodal practices and annotations. Their goal is to help showcase innovative work and trends undertaken in research labs and to finalize and disseminate current methods and practices in digital humanities research.

## 1 Introduction

CORLI (Corpus, Langues et Interactions: Corpus, Languages, and Interaction) is a French consortium of people involved in linguistic research and teaching. It is one of several such consortiums involved in digital humanities overseen by Huma-Num (www.huma-num.fr). Huma-Num, which stands for Humanités Numériques (Digital Humanities), was created to help specialists in the humanities to use new digital material and services.

CORLI was established in January 2016 and it is foreseen that the structure will run for another four years. It is built on previous consortia for linguistics, one of which was specialized in corpus of written

material and the other in oral or multimodal corpora.

The purpose of consortium is not defined by a call for projects or other requirements from higher level institutions. The various objectives of the consortium were set up by the linguistic community itself, within limits set out by the scientific committee of Huma-Num. The self-organizing nature of CORLI is well attuned to the tradition of the work in French linguistics. It offers two advantages. First, actual work is based on existing practices and research, and people involved in CORLI tend to be motivated as they are defending or showcasing their own methods and practices. Second, proposals related to formats, tools, and practices are well accepted as they represent mainstream tendencies.

## 2 Organisation

CORLI has a Comité de Pilotage (CP: steering committee) which is responsible for deciding which annual goals CORLI should set itself and handling its relationship with the parent organisation Huma-Num. Financial responsibility and management is performed by the Institut de Linguistique Française (ILF: Institute of French Linguistics, which is part of the national research council (CNRS) structure – see http://www.ilf.cnrs.fr/). The head of ILF is the official head of CORLI.

The CP is made of specialists of the field of linguistics which are involved in corpus linguistics. The members of the CP are also representatives for the laboratories to which they belong. The number of CP members is not set, but is about twenty. This makes them very representative of the field. Membership to the CP can be easily changed, according to the needs or contingencies of the CP members.

CORLI is also organized in thematic workgroups (GT: Groupes de Travail). Membership of the GT is open to anyone who is involved in linguistics and all meetings are public. Members of the GT can be active, in the sense that they work on organizing scientific events, producing documents or handling people that might be hired on specific projects. However, they may also be observers, in the sense that they participate in the discussions or provide their own experience to other members. This allows the consortium's work to be based on the real-life, current needs or knowledge of the larger scientific community that it represents.

## 3 Goals

CORLI has set itself several goals for 2017. These goals are mainly follow up goals related to work that was accomplished in previous years.

### 3.1 Finalization of corpora

The goal of this action is to help corpora that already exist but are not yet ready for to be deposited in a corpus repository for dissemination. This concerns old material which was gathered before the generalisation of digital practices but also concerns ongoing projects that might need technical help or guidance.

### 3.2 Technical courses and information

Regular sessions are organized every year to introduce people to digital practices (recording, data handling and corpus deposition) and tools (transcription, querying, corpus exploration and annotation). This is done according to which tools and knowledge are useful for corpus linguistics and for sharing knowledge.

### 3.3 Describing resources

The goal of this project is to develop a tool to help people describe the data that they make available to the community.

### 3.4 Evaluation of resources

The ubiquitous use and requirement of digital tools, resources, data, is very important in current scientific work and projects. This means that the cost and quality of the material created and made available to the community has to be evaluated, as this is the case for any other scientific production.

The goal of this action is to help define evaluation criteria for linguistic corpora. This is in keeping with the French tradition of evaluation of scientific research.

### 3.5 Workgroup 1: Exploration and formats

The goal of this workgroup is to advertise the most useful and efficient tools for creating and exploring all types of linguistic corpora. Another goal is to showcase good practices in the use of metadata and formats. When necessary, this workgroup participates in the creation of tools dedicated to conversion formats or metadata handling, and also in the definition of corpora formats and metadata. The workgroup works hand in hand with both linguists users and tool developers.

### 3.6 Workgroup 2: Multimodality and new form of communication

This workgroup is dedicated to the development of cutting-edge practices, either in the human interaction domain (including gesture, visual languages, co-verbal communication), or concerning computer-mediated communication and social media corpora. This research calls for the use of new specific tools and formats (or the extension of previous tools and formats).

### 3.7 Workgroup 3: Corpus deposition and evaluation

The Corpus deposition and evaluation workgroup is closely linked to the work completed in corpus finalization and the evaluation of resources. Specific information sessions are organized to share good practice guidelines with colleagues.

### 3.8 Workgroup 4: Juridical information

Awareness and adherence to juridical regulations is very important for data that contain property rights and that might contain private information. This workgroup has already produced information which are available to any researcher and is responsible for following up new regulations, and especially European regulations.

## 4 Relationship with CLARIN

It is very important to be open to international work, knowledge, competence and sharing. French linguistics have, in some fields, a tradition of prioritising the French language above other languages or other linguistic traditions. Indeed, some scientific fields (e.g. text statistics, linguistic domains dealing with interpretation) are well established in France and researchers mainly work on the French language and communicate in French. This is not the case for all colleagues and differs according to the subfields and the different traditions in linguistics.

French linguistics, tools and data, would certainly benefit from being included in CLARIN - and CLARIN would also benefit from French expertise. This would open opportunities for European collaboration and help researchers who are already involved in international projects.

A large part of the work already completed within CORLI is highly compatible with the type of work achieved in CLARIN. We have been working on the use of the TEI format for all types of linguistic data, for written, oral as well as CMC and multimodal formats. We are also working on improving the quality of metadata to describe the corpora and associated documents. These metadata will be in the TEI format and, thus, will facilitate its use and harvesting.

CORLI also has a close relationship with the current CLARIN centres in France (C-centres: Cocoon and SLRD; candidate B-centre: Ortolang).

## 5 Conclusion

The CORLI initiative has goals that are very much aligned with the objectives of CLARIN. The consortium is currently assessing the benefits of a full integration of France into CLARIN. Now, our short-term aim is to explain, as best as possible, to French researchers and users of language data how the integration into CLARIN could offer opportunities to them and their research projects, but also explain to CLARIN users in other countries which kind of material they might find in the French data that are currently available.

## References


Cocoon: https://centres.clarin.eu/centre/33

Michel Jacobson, Flora Badin, Séverine Guillaume. Cocoon une plateforme pour la conservation et la diffusion de ressources orales en sciences humaines et sociales. 8es Journées Internationales de Linguistique de Corpus, Sep 2015, Orléans, France. 2015, <http://jlc2015.sciencesconf.org/>. <halshs-01319600>

Ortolang: https://www.ortolang.fr/

Jean-Marie Pierrel. ORTOLANG 1 : une infrastructure de mutualisation de ressources linguistiques écrites et orales. Actes de TALN 2014, 2014, Marseille, France. 2014. <hal-01113961>

SLDR: https://centres.clarin.eu/centre/28

Bernard Bel, Médéric Gasquet-Cyrus. Interdisciplinarity and the sharing of oral data open new perspectives to field linguistics. Colloque de l'AFLS : Regards nouveaux sur les liens entre théories, méthodes et données en linguistique française, Sep 2011, Nancy, France. 2011, <http://www.atilf.fr/afls2011/>. <hal-01514704>