CORLI: A linguistic consortium for corpus, language, and interaction

Christophe Parisse
Modyco, Inserm,
University of Nanterre,
France
cparisse@parisnan-terre.fr

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

Ciara R. 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.jacob-son@gmail.com

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

Abstract

CORLI is a consortium of Huma-Num, an organization that helps to develop digital humanities in France and provide services for this. CORLI is a consortium dedicated to linguistics and includes all aspects of linguistic research and development.

CORLI has a key role in corpus linguistics in France, and it can act as an interface or a facilitator between CLARIN and the scientific community of linguists. As France just joined CLARIN as an observer, the role of the consortium CORLI is very important in organizing the relationship between CLARIN and the French community.

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 broadcasting guidelines related to research and technical practices, especially about linguistic corpora. Finally, CORLI organises 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

1.1 What is CORLI?

CORLI (*Corpus, Langues et Interactions*: Corpus, Languages, and Interaction[1]) is a French consortium of Linguistics laboratories, gathering people involved in linguistic research and teaching. It is one of several such consortia involved in digital humanities overseen by Huma-Num[2]. Huma-Num, which stands for *Humanités Numériques* (Digital Humanities), was created to help specialists in the humanities use new digital material and services.

CORLI is steered by a board of people and laboratories representing the linguistic research community in France. Although the consortium's specific goals and the set of people and laboratories involved in CORLI may change from one year to another, the general goal is to promote the creation and the use of corpora in linguistics, and to represent the whole research community.

CORLI, as a consortium, focuses on knowledge information. It provides help to researchers through information support. It is also involved in facilitating corpus creation and dissemination, and designing tools or formats to handle linguistic data. The technical support for linguistic data is handled by the French CLARIN centres (ORTOLANG[3], Cocoon[4], SLDR[5]) or in some cases by the Huma-Num technical support.

1.2 Huma-Num

The goal of Huma-Num is to help and promote the use of digital technology for all humani-ties. For this purpose, they developed both technological and human responses to the queries from researchers and users in the humanities. Technological responses are means to store, process, broadcast, disseminate, search, and archive data. Human responses are consortia that target specific fields of study (for example linguistics, music, ethnology, etc.), or specific digital material used in corpora and databases (for example, maps, pictures, 3D images, etc.). The goal of a consortium might vary from one to another, but their general purpose is to develop and increment the digital data available, and to provide information and requirement about good practices for digital information. In a certain way, Huma-Num reproduces, for a larger set of scientific fields but at a smaller scale, what can be found in the CLARIN centres. On the one hand, some centres provide technical support, while on the other hand some centres provide knowledge information. Huma-Num, with the help of the French B and C CLARIN centres, provides technical support. CORLI provides knowledge information.

1.3 History of CORLI

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, that ran from 2012 to 2015. The first one was *Corpus-écrits* (Research Infrastructure for Written corpora), which was specialized in corpora based on written material and the second one was *IRCOM* (Research Infrastructure for Oral and Multimodal Corpora) which was specialized in oral or multimodal corpora. The goal of the initial consortium projects, as defined by Huma-Num, was to help for the creation and deposition of corpora. At that time, the focus was on making previously existing corpus projects available that, previously, had never been made public, either due to the lack of access to a repository or due to a lack of technical information/expertise. Both consortia decided that they had, on the one hand, to provide technical or financial help to corpora that were not yet disseminated, and on the other hand, to provide good practices about norms, formats, rights, and dissemination.

This was decided after consulting the community thanks to the organisation of several general assemblies that were open to all colleagues who wanted to be involved in the creation or use of corpora. More recent decisions have been taken by the steering committee (see below) or after consultation with the workgroups. The fusion of the two previous consortia into a unique consortium has not changed the overall organisation and purpose of the consortium.

[1] https://corli.huma-num.fr/
[2] http:// www.huma-num.fr
[3] https://www.ortolang.fr
[4] https://cocoon.huma-num.fr
[5] http://sldr.org

2 Organisation

CORLI (see figure 1) has a *Comité de Pilotage* (CP: steering committee) which is responsible for deciding which annual goals CORLI should set itself and for handling its relationship with the parent organisation Huma-Num. Financial responsibility and management is under control from 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 composed of specialists in the field of linguistics who are involved in corpus linguistics. The members of the CP are also representatives for the local research laboratories to which they belong. The number of CP members is not set, but is around twenty. This makes them very representative of the field. CP membership can easily be changed, according to the needs or contingencies of the CP members.

CORLI is also organized in *Groupes de Travail* (GT: thematic workgroups). GT membership is open to anyone who is involved in linguistics and all meetings are public. Members of a 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.

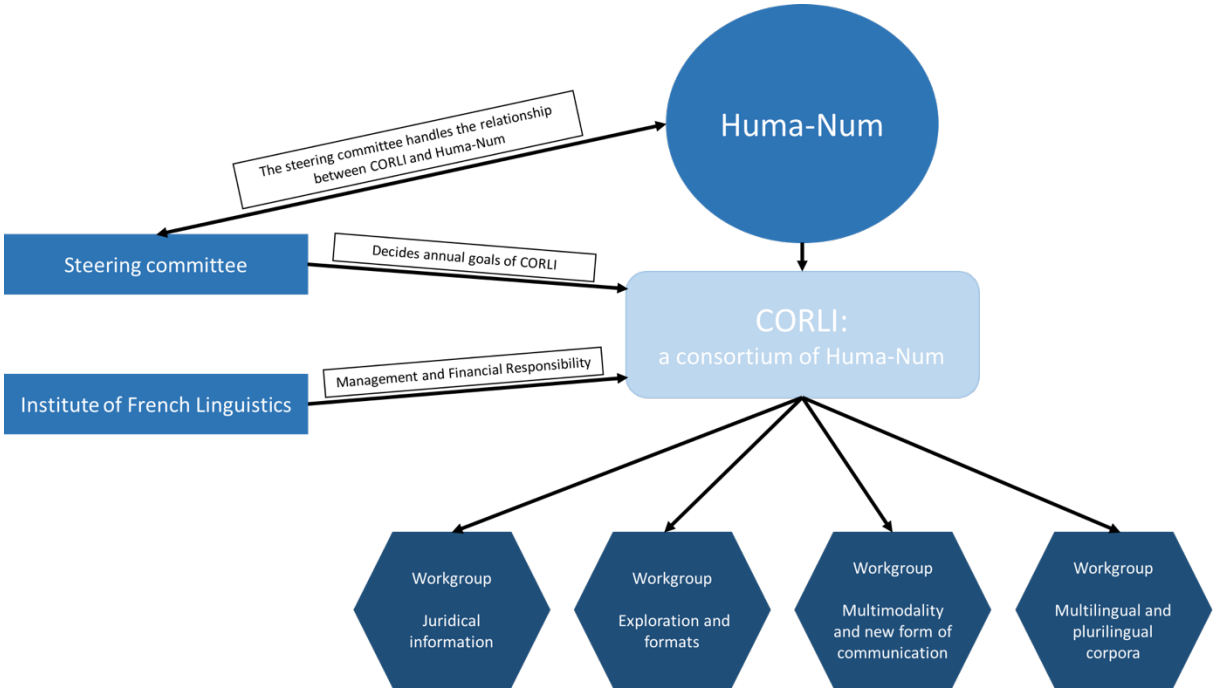

Figure 1: Organisation of CORLI

3 Goals

The goals of CORLI are defined each year by the CP. They are validated by the scientific committee of Huma-Num who sets the general goal of the consortia and is responsible for validating the project. The goals for 2017 are mainly follow-up tasks based on work that was accomplished in the previous years. The goals of CORLI are divided in two types of actions. General actions are under the direct control of the CP. They usually concern the whole community of linguistic research. The other types of action are the workgroups. They have a more specific purpose that either concerns only a part of the community or targets very specific tasks.

4 General actions

4.1 Describing resources

Although official repositories and archives are very efficient means to disseminate and preserve data about linguistic research or applied linguistics, they cannot and do not have to cover all uses and all types of corpora. Local research laboratories and projects are developing tools, creating data, managing working groups, and they cover a much larger range of formats, tools, and purposes than the official centres. All this material represents actual ongoing research. The nature of this material makes it difficult to be harvested using traditional open access inventories (OAI) because it is not yet normalized or made into standards. The purpose of this action is to provide a means to centralize information about these resources.

We are developing a portal that will make it possible for the people responsible for this type of resources to document themselves their product. It is necessary in this case to reverse the OAI mechanism, as it is only the people themselves who can provide information about their own resources, because this information is non-standard by nature (due to the on-going nature of the projects). The portal will make it possible to document the availability of digital linguistic resources, their formats and description, as well as how they can be accessed. This concerns a variety of data types including corpora, texts, lexicons, thesauri, dictionaries, etc. as well as tools such as software, libraries, scripts, stylesheets, query portals, etc.

4.2 Evaluation of resources

The large development of the use of corpora in linguistic research has led to an important investment of researchers and laboratories in the creation of corpora, but also digital tools, resources, and data. This takes a lot of time and resources, and is very important for the visibility of the people who worked for the projects. In a time where evaluation of research and justification of the means (financial or human) is very important, it is necessary to take into account the production of researchers and laboratories, not only in terms of paper or book publications, but also in terms of corpora or digital data.

This new development of the evaluation of research means that it is important in the evaluation to take into account the existence of digital material, but also the quality of this material, which is yet not clearly defined. The members of CORLI, as representatives of the research community, feel that this it is necessary to offer guidelines for corpus evaluation. The goal of this action is to help to define possible evaluation criteria for linguistic corpora, taking into account not only the size of corpora, but also their availability, the format, and the future uses by the community. The criteria would then be used for the evaluation of scientific work. The system of peer evaluation for corpora is in keeping with the French tradition of evaluation of scientific research.

4.3 Technical courses and information

The use and creation of corpora and digital data is possible only with the use of tools and methods that are rapidly evolving and developing. It is not possible for everyone to know which technology is the best or to learn to optimally use it by themselves. However, it is very important for senior or junior researchers to be up to date because it is necessary for their work and for teaching appropriate techniques and tools.

To answer this need, CORLI has been organizing annual technical courses. The nature of the courses is determined by the workgroups (see below) according to the actual needs of the community. For example, we have also included courses about data management, e.g. use of metadata or depositing corpora in repositories. The courses are divided into four main categories, as described below. Each category has several sub-categories. All courses are not presented each year, as this will depend on the actual requirements of the users.

Corpus annotation tools

The creation of oral or multimodal corpus linked with one or several media files requires the use of specialized software which is not easily mastered, especially at an advanced level. Several courses are

organized for different levels such as beginners, advanced, or experts for tools including CLAN[6], ELAN[7], Praat[8], SPPAS[9].

Corpus exploration tools

Existing corpora need to be exploited in the best way, including searching for data, classifying and categorizing, but also using advanced statistical tools and techniques. Some of these tools and methods have been developed in France. They are now frequently used in France by linguists and other researchers interested in text and corpus exploration. This explains why there is a large demand for training courses that cover these tools and methods. The tools that have been presented are, among others, TXM[10] (textometric exploration, R queries, interface with CQP), Iramuteq[11] (interface from texts to many R libraries), Unitex[12] (graphic interface for building text parsers), Le Trameur[13] (textometric exploration), Lexico5[14] (textometric exploration), DTM-VIC[15] (data and text mining), and Hyperbase web[16] (textometric exploration).

Video and sound recording

Recording audio or video data for corpus creation is a very time consuming activity. It is also expensive and cannot easily be done again and again if some hiccups appear during the data collection. For this reason, it is very important to acquire data of the best quality with minimal risk of failure. This technical course has been organized for several consecutive years with great success. Information was also provided about the format for best and most useful data compression, and also with regards to data mixing or movie/audio editing.

Metadata and corpus dissemination

There is a very large consensus about the use of metadata in order to deposit and to find corpora or any data in the most efficient way. However, creating correct and useful metadata is not an easy task. This is basically the work of specialists such as librarians or information specialists. Moreover, the access to such people is not always possible in small laboratories or for small projects. Also, metadata for objects such as linguistic corpora are different to metadata for other type of material. Thus, specialists cannot always help people in linguistics because metadata for linguistics is still a fairly new domain. So, it is important that people are able to produce, by themselves, the most adequate metadata, and that they are able to use this information to find data for themselves. This is why CORLI has offered information sessions on metadata use and creation.

Another requirement today, which is complementary to the creation of correct metadata, is to deposit corpora and tools in repositories such as CLARIN centres. This operation is not always so easily done because this does not simply correspond to copying the data that people have on their computers. The data is not always easily reused by other people and long term archiving is not always possible if the data is not in an open source format as is required by institutes like CINES[17] (long term digital archiving centre for France). CORLI has organized sessions to help people deposit their data, for example working interactively with their own data under supervision of a specialist in the field.

4.4 Finalization of corpora

Many corpora already exist but are not yet ready to be deposited in a corpus repository for dissemination, or this was never done due to lack of information or sometimes simply because of a lack of time and/or the opportunity to do this. CORLI, as well as the previous consortia *Corpus Ecrits* and *IRCOM*, but also projects such as ORTOLANG[18] (CLARIN B Centre candidate in France), have organized calls for

[6] http://alpha.talkbank.org/clan/
[7] https://tla.mpi.nl/tools/tla-tools/elan/
[8] http://www.fon.hum.uva.nl/praat/
[9] http://www.sppas.org/
[10] http://textometrie.ens-lyon.fr/
[11] http://www.iramuteq.org/
[12] http://unitexgramlab.org/fr
[13] http://www.tal.univ-paris3.fr/trameur/
[14] http://lexi-co.com/
[15] http://www.dtmvic.com/
[16] http://hyperbase.unice.fr/
[17] https://www.cines.fr/
[18] https://www.ortolang.fr/

finalizing corpora since 2013. The process was not exactly the same over the years, but most of the time this meant offering people the possibility to have some small financial or technical help to deposit their corpus in an official repository such the CLARIN B (candidate: ORTOLANG: Pierrel, 2014) and C (Cocoon[19]: Jacobson, Badin and Guillaume, 2015; SLDR[20]: Bel and Gasquet-Cyrus, 2011) centres in France. Some conditions need to be fulfilled to ask for this support:

- ✓ The corpus should be already advanced;
- ✓ The corpus should complement the already available corpora;
- ✓ The corpus should be open access for research;

For the year 2017, the maximal financial help for individual projects was 7000 €. There were 25 submissions to the 2017 call, which represented much more than the possible budget of 40 000 €. After the scientific evaluation of the different projects, 13 projects were accepted for 2017. The previous years had roughly the same amount of selected projects. The type of help requested in the submissions is always quite diverse, and can include for simple cases only information, or financial help for actual cleaning and depositing of the data, and in more complex cases coding of the data or finalisation of data collection so as to make it possible to deposit the data in its final form.

5 Workgroups

The most important creative work done in CORLI is the product of the workgroups. Workgroups target thematic subjects, which means that each of them brings together specialists in the domain. The principle that underlies most of the workgroups is that they are open groups for any person who has an interest in the theme of the workgroup. This can be people working on fundamental or applied research who are specialists of the field, or people that are not specialists but need to work on the subject. This way it is possible for a workgroup to know the specific needs of the people working effectively on the subject, and to have or build adequate responses thanks to the guidance of actual specialists with practical experience on the subject. The product of the workgroups can take the form of recommendations, of reference documentation, of norms or formats, or in some case of software development for small size projects.

5.1 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.

This workgroup was formed by merging two workgroups from the previous consortia; one that was working on written corpus exploration, and one that was working on designing a common format and methods that allow users to aggregate oral corpora. Initially, the existence of two workgroups was a consequence of the current state of research tools and corpora. For written data, corpora are less difficult to build, and so large corpora have existed for a long time, which called for the development of tools that were adapted to corpus exploration and statistical analysis. For oral data, corpora are difficult and expensive to build, and tools were first developed so that transcription and linking was easily done. Some tools exist for exploration of sound properties, but the tools that explore large oral language corpus are not in an advanced stage of development, if only because oral corpora (with included original media) are often small. Designing ways to use the same format for oral corpora, makes it easier now to build a large corpus. So this means that the tools made for written language become interesting to use with oral language corpora, which explains why both groups currently work together. Several actions are in progress in the workgroup. Results, whenever it was possible, have been presented at corpus linguistic conferences.

Exploration: Methods, tools and visualisations for analysing and processing corpora
The goal of this action is to find out what methods and tools exist for analysing and processing corpora, which formats they use, and how data can be prepared for this purpose. The format used by the tools will be taken into account by the other actions (see below) so that conversion between formats and

[19] https://cocoon.huma-num.fr
[20] http://sldr.org

description of metadata can directly target the tools that researchers use. The discussion about the actual use of the tools in the laboratories, which means how much it is used and how well mastered it is, provides information that is used to decide which technical courses and information (see above) is the most interesting to organize. This workgroup has led in the previous years to the publication of a book (Poudat and Landragin, 2017).

Formats and Metadata

Sharing corpora is highly dependent on two conditions: 1) using a common scheme for transcription and metadata format; 2) the quality of the information available in the metadata.

The workgroup has worked, for quite a few years, on using TEI as a support for oral language transcription and sharing. The work is based on the TEI Oral ISO format (International Organization for Standardization, 2016).

Good quality metadata must make it possible to describe the method used for the creation of the data and the content of the data. The basic level of metadata (Dublin-Core) used in the corpus repositories is often insufficient for fine-grained scientific purposes. This is not a question of format, or of the quality of the existing metadata. This is just because, in the linguistic data, a higher level of semantic content is required for research purposes.

Some projects do present more complex metadata, but when metadata go beyond Dublin-Core level, then the content might be different from one repository or one project to another. To avoid this and to encourage people to create fine-grained metadata, the workgroup has described a set of metadata for the analysis of oral language corpora that is considered as "minimal" in the sense that is contains enough information to create metadata of very good quality. The same work is planned to be done for written corpus metadata.

Tools for format and metadata

To make it possible for different users to use and produce the same metadata, it seems important for non-specialist users to have a tool available that is easy to use and that produces a format that can be automatically processed. We chose to develop a specific tool for this purpose whose settings can be customised and that is also easily accessible on the Internet. The tool produces a web interface in a web browser. This interface is easily changed with a configuration file. The result file is an XML file, which format is described in the configuration file. The tool is already available in its first version[21], but is still in the testing phase. The tool edits only the specific nodes that have metadata information in an XML file and leaves other data unchanged. It can, therefore, be used as a complement to other software programmes that edit XML files. The present format of the metadata is based on the TEI. Conversions to other metadata format such as CMDI will be done automatically.

Tools for format conversion

The existence of a common format for the transcription of oral language is interesting if it is easy to produce data in this format. For this purpose, a conversion tool has been developed. It allows an easy conversion from the TEI structure to the major oral language transcription formats used in France (CLAN, ELAN, Transcriber, Praat). The development of this tool was shared with ORTOLANG. It also allows users to convert back from the TEI to the other formats. No data is lost in the conversion to TEI and the conversion back to the same format. Some information can be lost when using the tool to convert between CLAN, ELAN, Transcriber and Praat formats, as these formats have different limitations in the nature of the data that they can store. We tried to keep the data lost in conversion between application formats as minimal as possible. The software is open source and freely available[22]. The use of the software to aggregate multiple corpora was presented by Parisse et al. (2017).

5.2 Workgroup 2: Multimodality and new form of communication

As for the previous workgroup, the multimodality workgroup brings together colleagues who were nvolved in the previous consortia. One consortium was working on written data and was taking into account new communication modes and especially computer-mediated communication (CMC), and one consortium was working on multimodal oral communication, including domains such as gestures (co-verbal gestures and sign languages). The new workgroup wishes to extend its target domain outside of

[21] http://ct3.ortolang.fr/teimeta/readme.php
[22] http://ct3.ortolang.fr/tei-corpo/

the field of linguistics, for example to domains such as education or sport sciences. The goal of the workgroup is to find common points and specificities of domains that link verbal and non-verbal data, and to propose solutions that are both useful and as generic as possible. Mixing communities that work on CMC data and sign language is one of means to reach this difficult goal.

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. In the human interaction domain, one goal is to integrate representations of new data types into corpora, such as motion capture, eye-tracking, EEG. Also, for sign language studies, a goal is to find representation systems that do not need the exclusive use of gloss in another language, but can represent movement of the hand and the body, for example. These new practices call for the organisation of dedicated training sessions that will be organized in the future.

The group will also follow up on the work on representation of CMC, network communication, and all type of hybrid communication. This will mix oral and written language representation, and comments which are found in a lot of collaborative systems (e.g. collaborative edition, wiki, online press, video, etc.).

All this work calls for harmonisation of the structure of the data that is used to create and deposit corpora. This is especially true because this type of data is new and changing a lot and very rapidly. Annotation proposals exist already for multimodal (TEI proposal of Oral data, TEI for Linguist SIG[23]) and for CMC (TEI for CMC SIG[24]). However, these proposals still need to be improved to take new developments into account. This will also represent a follow-up workshop with a similar theme organized previously by the consortium *Corpus-écrit*s. The current actions include mostly the production of good practice manuals for multimodal annotation and corpus creation. The workgroup is involved with current European research on CMC (see Beisswenger and Wigham, 2017), including participating in the annual CMC and Social Media Corpora for the Humanities conference series (cmc-corpora.org) and organising training sessions on structuring CMC corpora in TEI in association with CLARIN.

5.3 Workgroup 3: Multilingual and plurilingual corpora

This workgroup brings together researchers working on oral or written corpora of culture with a written tradition and researchers working on oral corpora of culture with oral tradition only. One goal of the workgroup is to share experience between the two communities, especially about the tools used for research and the theoretical aspects of the work. More specifically, the subjects under study are:

- ✓ Creation of oral or written corpora for language used by a whole country or a large community as opposed to creation of a corpus of a new language spoken by a small community: which are the best tools to be used, who are the best annotators, how can research be prioritized?
- ✓ Quantitative use of massive corpora of frequently studied languages vs. quantitative use of small corpus of unfrequently studied languages: which statistical models and methods can be used, which theoretical questions can be targeted?

The workgroup will organize training sessions for multilingual and plurilingual data, as well as training sessions for statistical processing. The group plans to promote the creation of multi-plurilingual corpora and to organize workshops on this subject. In 2017, a first panel was organized in Villejuif, France. Plans for the following years include the organisation of large size colloquium and the production of white papers on the subject. Also, the use of collaborative annotation with specialists of different languages is a promising option that needs to be developed.

5.4 Workgroup 4: Juridical information

Awareness and adherence to juridical regulations is very important for data that are subject to property rights and that might contain private or sensitive information. This workgroup has already produced white papers on the subject. These papers are freely available in the previous IRCOM website[25] and were produced in collaboration with other consortia. New development will be needed in 2018 to follow up new regulations, especially European regulations.

[23] http://www.tei-c.org/Activities/SIG/CMC/
[24] http://www.tei-c.org/Activities/SIG/CMC/
[25] http://ircom.huma-num.fr/site/p.php?p=groupetravail5

6 Relationship with CLARIN

In 2017, France joined CLARIN ERIC with an observer status. This was considered as a good opportunity to integrate the European research effort in making linguistic data freely available for everyone. French linguistics, tools, and data, would certainly benefit from being included in CLARIN - and CLARIN could also benefit from the 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. First of all, CORLI's main objective is to make all linguistic corpora in France available in one of the repositories that are already CLARIN centres, or are on the verge of becoming a CLARIN centre. The existence of CLARIN centres in France is not surprising because joining CLARIN was an old objective in France. So, although France is a recent CLARIN member and only an observer, many of the CLARIN principles were already effective in France.

The oldest centres are Cocoon and SLDR. Cocoon (https://cocoon.huma-num.fr) is a digital resource centre from and for the humanities and social sciences communities. The resources managed by this centre are speech recordings (audio or video) that are potentially accompanied by annotations and documentation. The services offered by the centre include storage, long-term preservation, integrity management, identification, description, curation and access to resources. The centre is based in Paris and is hosted by the CNRS/Huma-Num infrastructure. SLDR (http://sldr.org) is a centre that manages resources for spoken language and multimodal data. The centre covers storage, long-term preservation, and permanent identifiers. It is now integrated into the technical infrastructure of ORTOLANG, which is hosted at INIST[26]. ORTOLANG is a new French centre that aims to preserve and extend the work completed at SLDR and CNRTL[27]. Thus, the goal, as is the case for all French centres, is to cover storage, long-term preservation, and permanent identifiers. The data handled at ORTOLANG includes spoken and written language, as well as tools, lexicon, and terminologies. Any material that concerns language can potentially be preserved at ORTOLANG. The centre is based in Nancy, Aix-en-Provence, and Nanterre. The technical infrastructure is hosted by INIST in Nancy. All centres use OAI-PMH protocols, with metadata in OLAC and CMDI formats. ORTOLANG is harvested by the VLO of CLARIN.

Secondly, CORLI shares with other French initiative such as Isidore[28] from Huma-Num the belief that the quality of metadata is vital to the dissemination and use of language corpora. So CORLI has stressed many times how important metadata are and is working with the objective of improving the metadata.

Thirdly, good practices and sharing information are key to sharing and reusing data. This is why CORLI builds upon previous work that emphasizes the importance of sharing good practice guides, using well-known tools and metadata, and sharing information. The use of well-known formats (TEI, CMDI) is strongly encouraged.

Lastly, CORLI and a large part of the French community believe very strongly in open and free data, whenever this is possible. In most cases, people who deposit data use a CC-BY-NC licence, or another free access licence. This will make the data that is available in France also available to foreign partners.

CORLI has already established working relationships with foreign partners, for example regarding the work on oral transcription format and on CMC corpora. CORLI has a close working relationship with the already existing CLARIN centres in France (C-centres: Cocoon and SLRD; candidate B-centre: ORTOLANG). This relationship is strong through the calls for corpora finalization. For projects financed through these calls, all data must be deposited in the centres. It is also strong through the use of standard formats and metadata. For example, the metadata from the French centres is already harvested by the VLO of CLARIN. Whenever it is possible, the metadata format is CMDI. When it is not the case, a conversion to CMDI format could be organized.

7 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-

[26] http://www.inist.fr/
[27] http://www.cnrtl.fr/
[28] https://www.rechercheisidore.fr/

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 laboratories and projects, but also explain to CLARIN users in other countries what kind of material they might find in the French data that is currently available.

This has been done already twice. Once in a whole day session organized in Paris in September 2017 where researchers from other CLARIN members (Norway, Denmark, Germany, Italy) presented how CLARIN worked in their country and the implications for their own research. A second presentation was held in Montpellier in November 2017, with a conference and a poster presentation of CLARIN.

Our aim now is to organize ourselves in such a way as to be able to apply to become a CLARIN K Centre. It seems to us that the work we are currently accomplishing is very close to what a K Centre should do, and we feel like this application could be one of the ways to push France into becoming a full member of CLARIN. It would also help us to make the best use of the tools and services provided by CLARIN, and to ensure a productive dialog with the CLARIN community.

References

[Beisswenger Michael, Wigham Ciara. R., et al. 2017] Michael Beisswenger, Ciara. R. Wigham, et al. 2017. Connecting Resources: Which Issues Have to be Solved to Integrate CMC Corpora from Heterogeneous Sources and for Different Languages? In Stemle, E. and Wigham, C.R. (2017). (eds). *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities* (cmccorpora17). 3-4 October 2017, 52-55.

[Bel and Gasquet 2011] Bernard Bel, Médéric Gasquet-Cyrus. 2011. 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.

[International Organization for Standardization 2016] International Organization for Standardization. 2016. *Language resource management - Transcription of spoken language* (ISO/DIS Standard 24624) - https://www.iso.org/obp/ui/#!iso:std:37338:en

[Jacobson, Badin and Guillaume 2015] Michel Jacobson, Flora Badin, Séverine Guillaume. 2015. 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.sciences-conf.org/>.

[Parisse, Benzitoun, Etienne, Liégeois 2017] Christophe Parisse, Christophe Benzitoun, Carole Etienne, Loïc Liégeois. 2017. Agrégation automatisée de corpus de français parlé, *Journées de Linguistique de Corpus*, Grenoble, Juillet.

[Pierrel 2014] Jean-Marie Pierrel. 2014. ORTOLANG : une infrastructure de mutualisation de ressources linguistiques écrites et orales. *Actes de TALN 2014*, Marseille, France http://talnarchives.atala.org/TALN/TALN-2014/taln-2014-demo-001.pdf.

[Poudat and Landragin 2017] Céline Poudat, and Frédéric Landragin. 2017. *Explorer un corpus textuel : Méthodes - pratiques - outils.* De Boeck Supérieur, 240pp.