CORLI: The French Knowledge-Centre

Efstathia Soroli, Céline Poudat, Flora Badin, Antonio Balvet, Elisabeth Delais-Roussarie, Carole Etienne, Lydia-Mai Ho-Dac, Loïc Liégeois, Christophe Parisse

CORLI: The French Knowledge-Centre

Eva Soroli
University of Lille, CNRS, UMR 8163, MESHS, France
efstathia.soroli@univ-lille.fr

Céline Poudat
University Côte d'Azur, CNRS, BCL, France
celine.poudat@univ-cote-dazur.fr

Flora Badin
LLL-CNRS, Orléans, France
flora.badin@univ-orleans.fr

Antonio Balvet
University of Lille, CNRS, UMR 8163, MESHS, France
antonio.balvet@univ-lille.fr

Elisabeth Delais-Roussarie
University of Nantes, CNRS UMR 6310-LLING, France
elisabeth.delais-roussarie@univ-nantes.fr

Carole Etienne
ICAR-CNRS, Lyon, France
carole.etienne@ens-lyon.fr

Lydia-Mai Ho-Dac
CLLE, University of Toulouse CNRS, UT2J, France
hodac@univ-tlse2.fr

Loïc Liégeois
University of Paris, France
loic.liegeois@univ-paris-diderot.fr

Christophe Parisse
INSERM, CNRS-Paris, Nanterre University, France
cparisse@parisnanterre.fr

Abstract

As a first step towards increasing reproducibility of language data and promoting scientific synergies and transparency, CORLI (Corpus, Language and Interaction), a consortium involving members from more than 20 research labs and 15 Universities, part of the French large infrastructure Huma-Num, contributes to the European research infrastructure of CLARIN through the establishment of a knowledge sharing centre: the French Clarin CORpus Language and Interactions K-Centre (CORLI K-Centre). The purpose of the CORLI K-Centre is to provide expertise in corpus linguistics and French language, and support academic communities through actions towards FAIR and Open data. We describe the development of the CORLI K-Centre, its scope, targeted audiences, as well as its intuitive and interactive online platform which centralizes and offers both proactive and reactive services about: available language resources, databases and depositories, training opportunities, and best research practices (i.e., on legal/ethical issues, data collection, metadata standardization, anonymization, annotation and analysis guidelines, corpus exploration methods, format conversions and interoperability).

1 Introduction

More and more researchers underline the need to give data greater value, make them digital and interoperable as well as enhance their propensity for reuse. As a step towards increasing reproducibility of the data and promoting scientific collaboration and transparency, a group of researchers (Wilkinson et al. 2016) postulated the so-called FAIR principles (making data findable, accessible, interoperable and reusable). Such guiding principles are relevant for any scientific discipline but also increasingly relevant for linguistics, especially in the fields of corpus linguistics, natural language processing and computational linguistics typically characterized by great variability, isolated data collection practices, heterogeneous formats, etc. (Ciamiano et al. 2020).

In order to alleviate such issues related to incompatibility and promote interoperability, we propose to contribute to the European research infrastructure of CLARIN through the establishment of a French Clarin Knowledge Centre, the CORLI K-Centre, in the domain of corpus linguistics and French language based on the French CORLI (Corpus, Language, and Interaction) consortium (Parisse et al. 2018).

The aim of this paper is to share our experience and our network in developing a new K-Centre and to benefit from knowledge and recommendations from existing European K-Centres.

1.1 France, CLARIN, Huma-Num and the consortia

In the last decade, the French government and the CNRS (French National Research Institute) have funded several projects and initiatives to provide methods and tools to researchers working on digital humanities. Data centres dedicated to storage and dissemination of language data have been created, such as ORTOLANG and COCOON (both browsable centres in the CLARIN VLO), aiming at providing access to databases within sustainable digital platforms.

Huma-Num is currently the main player in the field. Supported by the CNRS and the French Ministry of Research, Huma-Num is a large research infrastructure with international reach devoted to Social Sciences and Humanities that promotes through technical support and funds research networks called *consortia*. CORLI is one of these consortia. Founded in 2016, CORLI provides services similar to those offered by CLARIN K-centres, and is dedicated to consensus-based recommendations and digital solutions in corpus linguistics.

1.2 The French CORLI consortium

CORLI is an open consortium of more than fifteen Universities, twenty laboratories and hundreds of researchers around France specialized in corpus linguistics and covering a very wide set of theoretical approaches (previously exposed by Parisse et al. 2018). CORLI's primary role is to be a reference centre for corpora and language resources, and a network that creates and distributes a wide array of language data, tools, methods and digital solutions. More specifically, we promote methodological approaches for corpora, from their collection to their storage, and we have become instrumental in promoting good practices, digital tools and standards by encouraging researchers to make their methods, outcomes and data open, accessible and reusable. As an incentive, we: (a) provide financial help to standardize and complete already existing corpora (oral, written, multimodal) through a yearly call for proposals; (b) closely monitor the development of new data; (c) support training sessions in digital annotation and analysis tools; and (d) we are attentive to the needs of the community, especially in domains such as metadata standardization, anonymization, corpus exploration guidelines, format conversions, interoperability as well as legal and ethical matters.

2 CORLI as a K-Centre

Although most of our goals are close to what other CLARIN K-Centres do, creating a new K-Centre calls for some clarification of our actions and specificities.

2.1 K-Centres and the CORLI K-Centre's expertise

As mentioned above, the *CORLI K-Centre* has two main areas of expertise: (a) corpus linguistics; and (b) the French language. With respect to corpus linguistics, and unlike other K-Centres specialized on specific topics such as computer-mediated communication, multimodality, sign language or learner corpora (i.e. the *K-Centre for Atypical Communication Expertise*), our centre's scope is closer to what more general centres do, such as the *CKLD: CLARIN Knowledge-Centre for linguistic diversity and language documentation* and *the Norwegian Centre for Research Data*, in that we offer expertise on data collection and data processing, as well as assistance for corpus building, annotation and data management.

With respect to the French language and its varieties, our Centre has similar goals to those of the *Czech CLARIN K-Centre for Corpus Linguistics*– structured around the Czech National Corpus and covering a wide range of actions (e.g., centralisation and mapping of language data and resources) – in that we offer centralized information about existing national corpora, annotation manuals and guidelines for the analysis of the French language and the languages of France to researchers working in this domain or interested in crosslinguistic comparisons.

The CORLI K-Centre is specialized, however not limited, to French language corpora, and thus can provide expertise on any type of language resources and language technology, as well as training opportunities for acquiring maximum autonomy in building and sharing language data – actions that dovetail with the objectives of other centres, such as the *CLARIN Knowledge Centre for South Slavic languages (CLASSLA)* and the *CLARIN K-Centre DANSK - DANish helpdeSK*.

2.2 Objectives and actions of the CORLI K-Centre

The main ambition of the CORLI K-Centre is to achieve a transformation of the research lifecycle in corpus linguistics and French language studies: offer expert advice from a panel of experienced

investigators about available databases and digital tools, provide resources to enhance the quality and reporting of linguistic and related research, support junior and early stage researchers in their training and development, and encourage FAIR data creation, edition and reuse.

In order to achieve these goals, the CORLI K-Centre functions as an intuitive and interactive on-line platform (already available from: https://corli.huma-num.fr/en/kcentre) which centralizes and provides cross-border access to knowledge through both proactive and reactive services.

With respect to proactive knowledge sharing, the CORLI K-Centre offers through a thematically organized website information about: (a) data sharing and access, (b) metadata standardization procedures, (c) format conversion and available software for language processing, (d) corpus exploration methods and tools, (e) guidelines and available manuals for corpus annotation, (f) legal issues related to corpus management and use and (g) training opportunities. The development of a FAQ (Frequently asked questions) page addressing common concerns in these topics (e.g. copyright, research ethics, research design, data collection, automatic analysis, corpus exploration methods) will further contribute to information access. The users of the CORLI K-Centre platform will have the possibility to access most knowledge through the website of the centre, and alternatively through the FAQ, where other landing pages will offer the possibility to redirect to related content (e.g., to ERIC, CLARIN, TalkBank, etc.) and thus continue the journey ideally without the need for outside assistance.

In cases of requests for further assistance, the CORLI K-Centre offers an additional reactive knowledge-sharing service established thanks to a pool of researchers and data specialists who can provide further information if needed. The way the users will interact with the webpage and the provided knowledge is of vital importance to the CORLI group. For this reason, a contact form has been integrated to the platform (easily accessible from a separate Contact-Us button) offering the possibility to the users who cannot find the answer to their questions to contact the centre directly.

3 Methodological challenges and solutions

Some of the greatest challenges in corpus linguistics are related to the great variability of practices as well as to the diversity of the corpora themselves (Cox 2011). With respect to the nature of the corpora, language data, by definition, present a huge variability: some researchers work with written data, others with spoken or gesture data; some focus on child language, others on adult use; some are interested in special populations and case studies (bilinguals/multilinguals, people with language disorders), others on natural language processing using very large corpora. With respect to the practices, investigators, according to their research questions and targeted populations, often opt for isolated data collection practices, incompatible annotation systems/formats and variable (in-house) management, storage and analysis methods that only rarely address ethical issues or allow for sharing and reuse. In addition, irrespective of types and formats and officially since 2018, researchers are invited to provide detailed information on the procedures for data collection, storage, protection, archiving and destruction of the collected data and thus conform to the General Data Protection Regulation (GDPR).[1]

As a first step towards increasing interoperability, transparency, protection and reproducibility of language data, CORLI has been managing, for several years now, six working groups that address these challenges. More specifically, the groups follow a committee approach, and produce consensus-based guidelines and recommendations in the following areas: Interoperability, query and annotation tools; Multimodality and new forms of communication; Multilingualism; Legal issues and Data protection; Best practices for corpus annotation; and Corpus assessments.[2]

The expertise gathered by the pool of specialists involved in the above groups has led to a great amount of outcomes, useful to linguists (all levels of academic expertise) but also to anyone working with corpora or interested in language use, databases, digital tools for data exploration and data management (engineers, data scientists, educators, etc.). Based on the outcomes of previous and current work, a series of knowledge sharing documents and tools are developed that can be used for serving the CLARIN community at large and thus meet the needs of a broader audience (e.g., Table 1 below).

[1] The GDPR is directly applicable in the EU Member States since 25 May 2018. The full text is available here

[2] For further information about the activities of our working groups see: https://corli.huma-num.fr/en/projectgroups

Resources	Purpose
Knowledge management diagrams	Data collection steps (video/audio data, constraints, to-do lists)
	The corpus lifecycle (including iterative processes, archiving and reuse)
	Flow diagram for corpus development and annotation methodology
Practice papers & Guidelines	Best practice recommendations for metadata, format conversions and sustainability
	Guidelines for anonymization and data protection
	Guidance to students, researchers, authors, editors and publishers about proper citation of language datasets, research projects, new annotations in archived corpora etc.
Technical manuals & useful documents	Templates of informed consents
	Manuals for transcription tasks with most commonly used software (Clan, Praat, Elan, Transcriber etc.) and minimal transcription recommendations
	Manuals: research methods and analysis (e.g., computer-mediated communication, sign language, multilingual corpora)
Bloopers	Typical errors in annotated corpora, in metadata etc.
	Most common anonymization issues
	Examples of typical speech errors in corpora (e.g., special populations, child data)

Table 1. Examples of useful knowledge sharing documents

The development of a K-Centre based on the CORLI consortium will bring acquired consensus-based expertise, services and digital solutions to a broader audience, will facilitate the creation of international synergies among actors interested in language corpora and French linguistics, and thus strengthen the participation of France (currently an observer) to the CLARIN infrastructure.

4 Conclusion

To conclude, the work of the groups on recommendations, good practices, tools and methods, as well as the activities and commitments of the network to help researchers facing new challenges (GDPR, open science, data management), make CORLI the appropriate organization for a CLARIN K-Centre. With expertise in corpus linguistics and French language, the CORLI K-Centre aims to become a major platform of knowledge sharing and communication among researchers and other actors interested in language and corpus linguistics (data scientists, engineers, educators, etc.). Built on the shared knowledge of the CORLI consortium, the new CORLI K-Centre will benefit from our past experience and current projects in tool development and practices, and contribute with actions that fit perfectly with the scope of other CLARIN K-Centres. The centre will provide through both proactive and reactive services useful information about available national and international databases and depositories, manuals and annotation techniques for French and other languages, corpus exploration methods, conversion platforms for interoperability, advice on legal issues, metadata standardization procedures, webinars and online training opportunities, and thus will contribute to international synergies and enhance knowledge and practice sharing.

References

Cimiano, Ph., Chiarcos, Ch., McCrae, J. & Gracia, J. 2020. *Linguistic Linked Data: Representation, Generation and Applications*. Springer International Publishing.

Cox, C. 2011. Corpus linguistics and language documentation: challenges for collaboration. In J. Newman, J., Baayen, H. & Rice S. (eds.) *Corpus-based studies in language use, language learning and language documentation* (p. 239-264). Brill, Rodopi.

Parisse, C., Poudat, C., Wigham, C. R., Jacobson, M., & Liégeois, L. 2018. CORLI: A linguistic consortium for corpus, language, and interaction. In *Selected papers from the CLARIN Annual Conference 2017*, Budapest, 18–20 September 2017 (p. 15-24). Linköping University Electronic Press, Linköpings universitet.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data 3*: 160018.