CLARIN K Centre: Development and Perspectives

Parisse Christophe, Poudat Céline

▶ **To cite this version:**

Parisse Christophe, Poudat Céline. CLARIN K Centre: Development and Perspectives. CLARIN Annual Conference 2023, CLARIN, Oct 2023, Leuven, Belgium. hal-04255184

HAL Id: hal-04255184
https://hal.parisnanterre.fr/hal-04255184v1

Submitted on 24 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

CORLI CLARIN K Centre: Development and Perspectives

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

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

Abstract

One of the primary objectives of the CORLI CLARIN K Centre is facilitate collaboration among researchers in the field of language sciences. The center aims to foster the development of projects that might be beyond the scope of individual researchers and to provide access to cutting-edge digital tools that enhance their scientific endeavors. These tasks are achieved by providing support and training in the utilization of modern digital tools designed for tasks such as corpus creation, annotation and data analysis. However, there are instances where the existing tools prove insufficient for the demands of language research. This can occur when these tools lack necessary functionalities, are unavailable in the required format or do not align with specific research needs. In light of these challenges, we will introduce two ongoing projects within CORLI that are focused on bridging the gap between researchers, technology and data:

- Open French Corpus: A centralized platform for accessing and utilizing existing corpora with shared tools.
- Collaborative annotation: use and improve existing tools; connect researchers, educators and students; develop a collaborative resource.

1 Introduction

The CORLI CLARIN K Centre[1] (Parisse et al., 2017; Soroli et al., 2020) was created in 2020. Comprised of members from over 20 French research labs and 15 Universities, the consortium is part of the lage French infrastructure Huma-Num. This infrastructure is dedicated to assisting researchers in the Humanities to use all types of digital data and tools. The CORLI CLARIN K Centre mainly aims to respond to users' needs regarding data and tools. We offer information, training and facilitate discussions and among academic users. These efforts aim to foster the development of projects and recommendations across various research areas involving language corpora.

It is during these panels and at our annual CORLI conference that proposals surfaced, addressing user needs in two domains of significant interest to the linguistic research community. In both instances, the projects CORLI undertakes build upon existing tools or standards that are already used and endorsed by the research community. However, these tools and standards fall short of fully meeting researchers' requirements. This gives rise to two main situations in which CORLI can play a pivotal role in assisting these situations to be resolved.

1) **Integration of tools and data**: Combining various tools or data components into a more comprehensive tool or dataset.

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

[1] https://www.clarin.eu/blog/tour-de-clarin-french-clarin-knowledge-centre-corli-corpora-languages-and-interaction

2) **Project expansion or adaptation**: Extending or modifying a research project to suit different scenarios or contexts.

2 An Open French Corpus

The initial proposal involved providing unified access to a high-quality corpus of the French language. Numerous independent corpora exist for French, often stemming from funded project. While these corpora might be substantial in size, the tools supporting their use might no longer be maintained once the project concludes. Additionally, issues with corpus format maintenance can hinder their usability. Alternately, some corpora result from collaborative efforts of researchers or laboratories over multiple years. While these corpora tend to exhibit exceptional quality, they might not attain the scale of those funded by larger projects. Nevertheless, they offer the advantage of being more recent, undergoing modifications and extensions. Diverse corpora are also present, such as those created for doctoral theses or cultivated by individual researchers through dedicated efforts.

In any case, these corpora are often clearly identified and accessible in well-known formats. Some are securely stored in institutional archives such as ORTOLANG or COCOON, university repositories, or even private repositories in some cases. In all cases, the corpora are accompanied by scientific publications describing their creation, format, and objectives.

CORLI's mission is to gather these scientifically validated sources into a single repository. Our objective is to standardize the format across all data and enable their utilization through a common set of tools. The efforts of CORLI are directed along three primary avenues:

1) **Harmonizing Metadata**: Establishing a fundamental metadata structure that can be universally applied for processing all data.
2) **Data Format Conversion**: Transforming data into both raw text and TEI (Text Encoding Initiative) formats, catering to distinct usage scenarios.
3) **Providing Processing Tools**: Furnishing tools capable of processing, querying, and displaying the complete dataset.

Whenever feasible, our approach entails automated conversion and processing. This approach not only facilitates the integration of future data—whether newly deposited into official repositories or updates to existing data—but also ensures efficiency in achieving our goals.

3 Collaborative annotation

Collaborative annotation plays a pivotal role within the linguistic community. The availability of extensive language corpora presents a challenge, as the task of editing and annotating an entire language dataset can be overwhelming for an individual. The process of corpus annotation demands significant time investment, ideally distributed across multiple contributors to make it feasible. While projects with substantial financial support can manage this, individual researchers, groups, or even entire laboratories face constraints.

Even in cases where annotation is performed automatically, there remains a need for manual oversight to verify and analyze the output of automatic processing. As a result, collaborative annotation has become a necessity in numerous instances.

Collaboration can take different forms. It may involve multiple skilled users, where the collaborative tools primarily facilitate data sharing and prevent redundant annotations. Alternatively, collaboration can include less experienced users. Here, the collaborative tools must enable researchers to compare several annotations and to resolve inconsistencies through an adjudication process.

The effectiveness of collaborative annotation is also influenced by the data format. While original data may be in text or TEI formats, tools already exist for editing such data, and CORLI will endeavor to leverage these existing resources. More intricate situations arise when the original data takes the form of images (such as handwritten materials or low-quality documents). In such cases, the image must be displayed, and annotations need to be linked to specific portions of the image. Lastly, annotations could also be generated automatically and subsequently checked manually.

The range of scenarios requires tailored annotation environments. Recognizing the impossibility of a one-size-fits-all solution, our approach is to focus on enhancing and utilizing three distinct tools that cater to the specific requirements of CORLI-affiliated laboratories:

1. TACT (Transcription and Annotation Collaborative Tool): The TACT initiative (https://tact.demarre-shs.fr/) centers around a web platform designed for transcribing and annotating text corpora. Our objective is to enrich the platform's capabilities, enabling it to facilitate crowdsourced transcription of parliamentary data.
2. INCEption (Integrated Corpus Exploration): INCEption (https://inception-project.github.io/) is a web-based annotation tool for corpus data. Our aim here is to improve the conversion of external data into the tool's internal format, encompassing data with pre-processed syntactic or semantic annotations. To this end, we've developed a converter (https://corli.huma-num.fr/convinception/#/sax2) that allows users to import and export XML corpora to and from Inception.
3. GUM (Grammatical Universal Dependencies Multilayer Corpus): The objectives of the CORLI-GUM are twofold. Firstly, we seek to foster collaboration among researchers, teachers, and Master's students. Researchers often lack annotators for their projects, while teachers express interest in engaging students in ongoing annotation initiatives. This approach enhances students' learning experiences and motivation. Secondly, we endeavor to develop an open-source, multi-layer corpus of richly annotated texts (https://gucorpling.org/gum/), following the example set by Amir Zeldes in 2017. The resulting resource will be made accessible to the wider community.

4 Conclusion

The ongoing objective of the CORLI CLARIN K Centre is to continue assisting researchers in France by providing access to the most effective linguistic tools, including those furnished by the CLARIN infrastructure. Simultaneously, we persist in the development, enhancement, and utilization of tools that researchers and laboratories within the CORLI network require for their work.

In alignment with our previous approach, we remain committed to augmenting existing tools rather than starting from scratch. This methodology not only proves more cost-effective but also resonates better with the community, as it aligns with their current practices and needs.

The CORLI CLARIN K Centre's operations are made possible through funding from the CORLI consortium, facilitated by a Huma-Num grant (https://www.huma-num.fr/les-consortiums-hn/), extending until 2024. This grant empowers us to continue our mission in supporting linguistic research and tool development within the community.

References

Christophe Parisse, Céline Poudat, Ciara Wigham, Michel Jacobson, Loïc Liégeois. CORLI: A Linguistic Consortium for Corpus, Language and Interaction. CLARIN Annual Conference 2017, Sep 2017, Budapest, Hungary. ⟨halshs-01636943⟩

Efstathia Soroli, Céline Poudat, Flora Badin, Antonio Balvet, Elisabeth Delais-Roussarie, et al.. CORLI: The French Knowledge-Centre. CLARIN Annual Conference 2020, Oct 2020, Barcelone (virtual), Spain. ⟨hal-03091629⟩

Zeldes Amir. The GUM corpus: creating multilayer resources in the classroom. Language Resources and Evaluation 51, 581-612, Springer.

Annex: Non-exhaustive list of corpora to be included in the OFC

- Written language corpora
  - Scientext https://scientext.hypotheses.org/corpus
  - Scienquest https://corpora.aiakide.net/

    - Archives parlementaires https://archives-parlementaires.persee.fr/
    - Consortium CAHIER https://cahier.hypotheses.org/
    - E-CALM https://www.ortolang.fr/market/corpora/e-calm
    - Corpus 14 https://www.univ-montp3.fr/corpus14/
    - Democrat https://hdl.handle.net/11403/democrat
- Spoken language corpora
    - CFPP2000 http://cfpp2000.univ-paris3.fr/search.html
    - ESLO http://eslo.huma-num.fr/index.php/pagecorpus/pageaccescorpus
    - CHILDES https://talkbank.org/DB/
    - CT3-ORTOLANG https://ct3xq.ortolang.fr/ct3xq/interro
    - PFC https://public.projet-pfc.net/transcription/
- Multiple-format corpora
    - CEFC-Orféo https://orfeo.ortolang.fr/, http://orfeo.grew.fr/
    - CoMeRe http://hdl.handle.net/11403/comere