Research infrastructures and FAIR principles in linguistics:
interdisciplinary perspectives, theoretical implications and practical applications

Convenor's details
Name: Efstathia SOROLI
Email: efstathia.soroli@univ-lille.fr
Affiliation: University of Lille & UMR 8163 STL lab, France
Webpage: https://pro.univ-lille.fr/efstathia-soroli

Name: Christophe PARISSE
Email: cparisse@parisnanterre.fr
Affiliation: INSERM, UMR 7114 Modyco & U. Paris Nanterre, France
Webpage: https://cv.hal.science/christophe-parisse

Description of the Workshop

Target Topics

Corpus linguistics, Digital humanities, Research infrastructures, Parallel corpora, Comparable learner corpora, Comparative Linguistics, Second Language Acquisition and Teaching, Historical Linguistics, Comparative Anthropological research, Computational approaches to discourse analysis, FAIR principles

State of the art

Digital technologies and information systems are now ubiquitous in Human and Social Sciences (HSS) impacting research, education and nearly every form of expression and creation [1]. Digital humanities –a discipline at the intersection of Computer science and traditional HSS (history, philosophy, linguistics, literature, anthropology, etc.) often considered as a “trans-discipline” that uses tools, digital methods, devices and heuristics borrowed from information science [2]– faces more and more challenges related to the increasing volume of data, corpora, applications, software, and to their great heterogeneity [3]. In this context, where great amounts of data become available in various formats and open science practices are generalized, the role of infrastructures becomes central in data sharing and standardization of practices for every interdisciplinary endeavor [4].

In the field of Language Sciences, this digital and interdisciplinary aspect is highly prominent and inherently integrated into methods and practices used in Corpus linguistics, Computational linguistics, Natural language processing and Experimental linguistics. Such a digital approach, although initially focused on large corpora of digitized texts focused on fine-grained quantitative analyses on word occurrences, calculation of frequencies of specific linguistic phenomena, and the typology of literary styles, now extends to other areas such as discourse analysis, textometry, modeling, and corpus-based teaching [5-7], enriched by techniques inspired by data mining, experimental simulation, machine learning and deep neural networks [8].

Objectives

The aim of this Workshop is to discuss the challenges associated with the surge in the volume of linguistic data, the diversification of approaches in this domain, and the pressing need for sharing knowledge, practices and resources. To address these challenges effectively, it is crucial to avoid duplication of efforts by adhering to common research infrastructures and to FAIR principles, thereby ensuring that linguistic resources are findable, accessible, interoperable and reusable [9]. While these principles are universally applicable across scientific disciplines, their significance has become increasingly relevant for Linguistics – a discipline characterized by a high variability in data collection practices and important heterogeneity in the formats used [10].

(Sub)disciplines involved and scope of the workshop

More specifically, during this Workshop, we will present the most important research infrastructures (national and European) that support the work of linguists through platforms offering access to data, advanced tools for their annotation, exploration, comparison and analysis, such as Ortolang, Corli, Humanum, RnMSH, Dariah and CLARIN [e.g., 11-14]. The first part of the Workshop will focus on those research infrastructures and support centers which actively contribute to the development of precious resources and the sharing of knowledge, not only for linguists but also for anyone interested in such resources (researchers, engineers, speech and language therapists, educators).

The second part of the Workshop will be devoted to case studies, tutorials, and examples of data and tool usage (e.g., use of parallel and comparable corpora) in specific research areas (e.g., study of L2 acquisition, typology research, historical linguistics, teaching, etc.). The goal is to provide hands-on demonstrations showcasing the diverse applications and wide-ranging theoretical implications of these resources and methodologies. Discussions will include proposed solutions to hypothesis testing, and to challenges in maintaining consistent coding principles, in utilizing common technologies and in standardizing sharing practices for improved replicability, interoperability, and collaborative efforts [15].

The final portion of the day (through registration only and limited to a maximum of 10 participants) will focus on a tutorial introducing basic tools for spoken discourse analysis using Talkbank and the CLAN/Computerized Language Analysis software [16].

Bibliography

[1] Clement, T.E. & Carter, D. (2017), Connecting theory and practice in digital humanities information work. Journal of the Association for Information Science and Technology, 68: 1385-1396.

[2] Mounier, P. (2010). Manifeste des Digital Humanities. Journal des anthropologues, 122-123 | 447-452.

[3] Burnard, L. (2012). Du literary and linguistic computing aux digital humanities : retour sur 40 ans de relations entre sciences humaines et informatique In : Read/Write Book 2 : Une introduction aux humanités numériques [en ligne]. Marseille : OpenEdition Press (généré le 03 mars 2022). Disponible sur Internet : <http://books.openedition.org/oep/242>.

[4] Joffres, A., Priddy, M., Morselli, F., Lebarbé, Th., Granier, X., Bertrand, P., Rodier, P., Melka, F., Camlot, J., Sinclair, S., Fatiha, I., Abela, C., Chayani, M., Parisse, Ch., poudat, C., Ginouvès, V., Sinatra, M., et al. (2019). “Building community” at the national and/or international level in the context of the Digital Humanities. Digital Humanities Conference, 2019, Utrecht, The Netherlands.

[5] Meunier, J. G. (2020). La rencontre du sémiotique et du «numérique» : Le rôle d’une modélisation conceptuelle. Semiotica, 2020 (234), 177-198.

[6] Longhi J. (2020). Proposals for a Discourse Analysis Practice Integrated into Digital Humanities: Theoretical Issues, Practical Applications, and Methodological Consequences. Languages 5(1): 5.

[7] Boulton, A., & Landure, C. (2016). Using corpora in language teaching, learning and use. Research and Teaching Languages for Specific Purposes, 35(2). https://doi.org/10.4000/apliut.5433

[8] Suissa, O., Elmalech, A., & Zhitomirsky-Geffet, M. (2021). Text analysis using deep neural networks in digital humanities and information science. Journal of the Association for Information Science and Technology, 73(2): 268– 287.

[9] 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

[10] Cimiano, Ph., Chiarcos, Ch., McCrae, J. & Gracia, J. (2020). Linguistic Linked Data: Representation, Generation and Applications. Springer International Publishing.

[11] Erhard, H. & Krauwer, S. (2014). The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), May 2014, 1525–31.

[12] Olivier Baude, Adeline Joffres, Nicolas Larrousse, Stéphane Pouyllau. Huma-Num : Une infrastructure française pour les Sciences Humaines et Sociales. Stratégie, organisation et fonctionnement. DH 2017, Aug 2017, Montréal, Canada.

[13] Parisse, Ch. & Poudat, C. (2023). CORLI CLARIN K Centre: development and perspectives. CLARIN Annual Conference 2023, 16-18 octobre 2023, Leuven, Belgique.

[14] Soroli, E. (2021). CORLI, the French Knowledge Centre for Corpora, Languages and Interaction. In Lenardic, J., F. Frontini & D. Fiser (eds.) Tour de CLARIN, volume IV: 40-45.

[15] Branco, A., Calzolari, N., & Choukri, K. (2018). LREC 2018 workshop proceedings: 4REAL 2018 Workshop on Replicability and Reproducibility of Research Results in Science and Technology of Language, 12 May 2018, Miyazaki, Japan. European Language Resources Association.

[16] Mac Whinney, B. & Fromm, D. (2022) Language Sample Analysis with Talkbank: An update and review. Frontiers in Communication, 7: 865498.

Overview of the planned slots (provisional)

Slot 1: Efstathia Soroli & Francesca Frontini (CLARIN-ERIC)	confirmed
Slot 2: Christophe Parisse (CORLI, Ortolang)	confirmed
Slot 3: Olivier Baude & Nicolas Larousse (HumaNum)	confirmed
Slot 4: Presentation of the RnMSH National Network	to be confirmed
Slot 5: Edward Gray (DARIAH)	confirmed
Slot 6: Thomas Gaillat et al. (Univ. of Rennes)	confirmed
Slot 7: Antonio Balvet (Univ. of Lille & STL)	confirmed
Slot 8: Annemarie Verkerk et al. (Univ. of Saarland, Germany)	confirmed
Slot 9: Matthias Urban (Univ. of Tubingen & DDL)	confirmed

Tutorial: Basic tools for spoken discourse analysis with TalkBank and CLAN (Computerized Language Analysis)

Provisional abstracts

Slot 1: Efstathia Soroli (University of Lille, STL, CLARIN) & Francesca Frontini (CLARIN-ERIC)

Title: CLARIN The European research infrastructure for linguistic data resources and technology

CLARIN, the European research infrastructure for linguistic data resources and technology, established in 2012, provides researchers with a platform enabling easy and sustainable access to language data and processing tools. CLARIN offers advanced services for corpus compilation, comparison, annotation, and analysis, facilitates cross-linguistic studies, supports format interoperability, and aims to enhance the potential for comparative research following FAIR principles through a distributed network of over 70 data repositories and knowledge centers in 25 countries. With the development of this international network and the offer of a secure, unified interface, CLARIN ensures continuous resource sharing among all stakeholders involved in building, exploiting, and using language corpora. Its mission is to ensure that knowledge and expertise are not fragmented but organized and accessible to anyone interested in such resources (researchers, engineers, educators). This presentation will cover: (a) how a dataset can be accessed/constructed/shared through the CLARIN platform (cf. Language families, the Virtual language observatory services) (b) the ways it can be cited (e.g., through the Virtual collections tool); (c) used/reused (e.g., through the Language resource inventory); and (d) processed (cf. Switchboard tools). An illustrative example will be provided based on the ParlaMint project, which provides parliamentary corpora in 17 languages. These corpora are designed with the same encoding standards (TEI ParlaMint) and can be utilized with the same tools and interface for comparative analysis. We will demonstrate that the value of this approach lies in sharing various language datasets, linguistic tools and collections. The goal is to enable researchers and anyone interested in these resources to access a comprehensive range of services, to navigate between databases and tools, while maintaining consistent encoding standards compatible with open science practices and FAIR principles.

Slot 2: Christophe Parisse (University of Paris Nanterre, Modyco, Corli & Ortolang)

Title: CORLI and ORTOLANG: Providing tools and guidance for sharing and reusing open science language data

Language sciences have a long tradition in data sharing at least for two main reasons: data collection is a costly procedure, and each language production is unique and non-repeatable. Therefore, recording and preserving language datasets is essential, especially for linguists. While language corpora have existed for more than 40 years, in France, language data were not consistently centralized, leading to great diversity of formats. This has changed in the past decade thanks to the CORLI initiative and infrastructures such as ORTOLANG. The goal of the CORLI network is to help researchers to use common formats and standards, and to save data in publicly available infrastructures such as ORTOLANG. The goal of ORTOLANG is to make data deposit easy, to support data management and processing, to preserve data in the short and long terms, and thus facilitate dissemination. This paper presents the work done on spoken language corpora (similar work is ongoing for written language in CORLI and ORTOLANG). Creating and disseminating spoken language data involves three main phases and the main aim is to help the researcher to handle these phases as efficiently as possible. The first phase involves corpus constitution, which includes collecting, formatting, and describing the data. This process typically involves utilizing specialized tools, the output of which is then converted in a standardized TEI for Spoken Language format. The second phase is depositing and providing adequate metadata – e.g., through the TEIMETA tool developed by CORLI. The third phase involves using or reusing the data for research purposes. This is usually done by using specific tools such as TXM (a textometric tool), any specific tools using R or Python, or more simple tools that make it easy to browse the data, such as the export feature of our TEI tool, TEICORPO, which produces adequate formats enhancing interoperability for language data.

Slot 3: Olivier Baude & Nicolas Larousse (CNRS, Huma-Num, UAR 3598).

Title: The role of the Huma-Num infrastructure in supporting language sciences

Huma-Num IR* is a French national infrastructure dedicated to supporting research project in SSH (Social Science and Humanities) in accordance with the principles of Open Science. More particularly for language sciences, this support takes several forms: in addition to the standard services offered by Huma-Num tailored to each step of the research data lifecycle (https://documentation.huma-num.fr/media/services-HN-en.png), some specific support is provided for this community at very different levels. Here are a few examples. At the core infrastructure level, Huma-Num is providing specific tools such as automatic speech recognition (ASR) system. At the national level, Huma-Num has labelled and financed the CORLI Consortium, which enables targeted actions like identifying and developing tools, providing guidance for good practices, organizing trainings etc. Huma-Num also supports the development of national repositories dedicated to language resources, ORTOLANG and COCOON, in particular by offering its long-term preservation service (https://documentation.huma-num.fr/en/partenariat-hn-cines-en) for the data hosted. At the European level, Huma-Num rely on its international network and its experience of the European ecosystem in order to play the role of facilitator, notably with the European CLARIN infrastructure.

Slot 4: (RnMSH)
Presentation of the RnMSH national network
To be confirmed

Slot 5: Edward Gray (Dariah)
Presentation of the Dariah international network
Title to be announced

Slot 6: Cyriel Mallart (LIDILE, Univ. de Rennes), Andrew Simpkin (NUI Galway, National Univ. of Ireland), Rémi Venant (LIUM, Univ. du Mans), Nicolas Ballier (CLILLAC-ARP, Univ. de Paris), Bernardo Stearns (Insight Centre for Data Analytics, Galway), Jen-Yu Li & Thomas Gaillat (LIDILE, Univ. De Rennes).
Title: Linguistic interoperability within a unified architecture: the case of Analytics for Language Learning

Modern approaches to quantitative linguistics rely on large datasets which are representations of linguistic observations made up of features of various dimensions. Analyses rely on such data representations making essential the principles of their construction. In the case of quantitative research methods, datasets are built automatically. This includes the sequential use of various types of NLP tools/components of software architectures that ensure data interoperability (Rehm et al., 2020; Sérasset et al., 2009). In this paper, we present the data architecture used in the Analytics for Language Learning (A4LL) project (Gaillat, 2022a) whose aim is to produce linguistic analytics for foreign language teachers. The system transforms learner writings into linguistic feature sets which are used to produce linguistic indicators of L2 writing performance. The system relies on an architecture composed of NLP tools. Multilingual Linguistic annotation is provided by UDPipe (Straka et al., 2016) and Stanza (Qi et al., 2020) and stored in a database. Several tools exploit this annotation to compute textual measures of different linguistic dimensions. Syntactic complexity ratios are extracted with the use of TAASSC (Kyle, 2016). Text cohesion ratios are computed with TAACO (Crossley et al., 2016). Noun-Verb collocations are extracted (Li et al., 2022) and their association strength is computed by NLTK (Bird & Loper, 2004), yielding a number of scores. Specific linguistic structural complexity indicators are computed on by language models predicting small groups of words related to each other by the same paradigmatic relationships (Gaillat, 2022b). Some lexical sophistication indicators are computed via BERT-trained language models (Devlin et al., 2019). All these indicators are filtered out as part of a graphical visualisation module of the architecture relying on interoperable data representations. The architecture is hosted on a virtual machine of the Huma-Num consortium. The Docker technology supports a modular implementation of the system in which all modules act as standalone programmes with the aforementioned roles. These programmes are related to each other thanks to APIs and a data transfer queuing system. Automation and interoperability are discussed as the cornerstones of quantitative data processing.

Slot 7: Antonio Balvet (STL, University of Lille, France)
Title: Teaching linguistics with Clarin resources: preliminary results for syntax

Slot 8: Annemarie Verkerk, Luigi Talamo & Andrew Dyer (Univ. of Saarland, Germany)
Title: Compiling and annotating mini-CIEP+: a sharable parallel corpus of prose

Slot 9: Matthias Urban (Univ. of Tübingen, Germany & DDL, CNRS)
Title: Using cross-linguistic data formats (CLDF) databases for historical linguistics and comparative anthropological research: the example of CINWA, a database from South American Indigenous communities

Slot 10: Efstathia Soroli (Univ. of Lille & STL UMR 8163 CNRS)
Title: Basic tools tutorial for spoken discourse analysis with TalkBank and CLAN (Computerized Language Analysis)

This tutorial introduces the CLARIN K-centre Talkbank and the CLAN program, designed within this project for linguistic analysis, and will cover three main parts: transcription, linguistic (semantic and morphosyntactic) annotation, and automatic analysis. Participants will learn CLAN's utility (e.g., for studying discourse, language learning, disorders). The program facilitates transcription, media linking, annotation and aids researchers and clinicians in hypothesis testing by computing indices like MLU, TTR, and IPSyn (among others). The examples provided will focus on child language data but are adaptable to other language (re)appropriation types like second language acquisition, acquired and developmental language disorders. The tutorial aims to empower young researchers with practical skills for their language studies.

NB: Tutorial limited to a maximum of 10 participants (registration required via email: efstathia.soroli@univ-lille.fr) For the registration, please provide the following: Name, Institution, your field/discipline and your research experience level: e.g., **Early-Stage Researcher** (ESR) in the first 4 years of their research careers from the date they embarked on a doctorate; **Experienced researcher** (ES) in possession of a doctoral degree and within their first 5 years of their career.