Workshop & Tutorial

Research infrastructures and FAIR principles in linguistics

Interdisciplinary perspectives, theoretical implications and practical applications

Organised by

Eva SOROLI, Associate Professor
efstathia.soroli@univ-lille.fr
University of Lille & UMR 8163 STL lab, France
Webpage: https://pro.univ-lille.fr/efstathia-soroli

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

The event is supported by the European Research Infrastructure CLARIN.

Description of the workshop

Target Topics

Research infrastructures, FAIR principles, Corpus linguistics, Crosslinguistic studies, Comparative Linguistics, Second Language Acquisition and Teaching, Historical Linguistics, Comparative Anthropological research, Discourse analysis, Resources, Tools.

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 working in Human and Social sciences, 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 database 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 digital 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 (upon registration) will focus on a tutorial introducing basic tools for spoken discourse analysis using the CLARIN Talkbank knowledge Centre and the CLAN software [16].

References

[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

Introduction
Slot 1: Eva Soroli & Francesca Frontini (CLARIN-ERIC)
Slot 2: Olivier Baude & Nicolas Larousse (HumaNum)
Coffee break
Slot 3: Chiara Chelini & Cécile Fabre (RnMSH)
Slot 4: Edward Gray (DARIAH-ERIC)
Slot 5: Christophe Parisse (CORLI, Ortolang)
Lunch break
Slot 6: Annemarie Verkerk et al. (Univ. of Saarland, Germany)
Slot 7: Matthias Urban (Univ. of Tubingen & DDL, Lyon)
Slot 8: Thomas Gaillat et al. (Univ. of Rennes)
Slot 9: Antonio Balvet (Univ. of Lille & STL)
Coffee break
Slot 10: Christophe Parisse, Alina Tsikulina, Adélie Buclier & Eva Soroli
Tutorial Session on discourse analysis with the CLAN (Computerized Language Analysis) software.
(registration required)
Cocktails

Abstracts and useful information

Efstathia Soroli (University of Lille, STL, CLARIN) & Francesca Frontini (CLARIN-ERIC)
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 participating 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 (e.g., via the Language families and 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 (see Switchboard tools). An illustrative example will be provided based on the ParlaMint project, supported by CLARIN, 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 seamlessly between databases and tools, while maintaining consistent encoding standards, employing common language technologies, and sharing a unified commitment compatible with open science practices and FAIR principles.

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

Chiara Chelini (RnMSH, UAR 3603) et Cécile Fabre (MSHS-T, UAR 3414)
The French national networks of humanities and social sciences centres, their platforms and the role of linguistics.

The national network of humanities and social sciences centres (*Réseau national des maisons des sciences de l'homme* - RnMSH) is a French infrastructure, a scientific cooperation consortium and a unit of the National Centre of Scientific Research (CNRS). The concept of « maison » (house) in the French name is difficult to translate into English and shows the idea of providing mutualised services, platforms and practices to the laboratories affiliated to the « houses ». The network federates 22 *maisons* all over the French territory which cover about the 70% of social sciences and humanities laboratories, in all disciplines. The network, managed by a collegial direction, is a place where collective intelligence and best practices can spread. Several interdisciplinary working groups and innovative actions are boosted by the network for the *maisons*, among which five networks of technological platforms, which, depending on their specific applications, are called: audio-visio, cogito, data, scripto, spatio. In 2023 the RnMSH launched a label campaign and 53 platforms have been labelled in 20 MSH, according to specific criteria. Implications of linguistics is mainly found inside cogito and scripto platforms, for instance in experimental research testing the cognitive bases of language acquisition and pathologies, and in digitalisation and textual analysis of different kinds of texts (modern or ancient) conducted with digital humanities methods. The 22 *maisons* are also an effective relay of the Very Large Research Infrastructures devoted to Social Sciences and Humanities and supported by the CNRS. Within this panel we would like to present a broad overview of the RnMSH missions and actions, and how the infrastructure can be useful to researchers in linguistics.

Edward J. Gray (Dariah-EU infrastructure)
DARIAH: Helping humanities researchers - including linguists - confront the digital turn

DARIAH-EU, founded in 2014, is European Research Infrastructure Consortium (ERIC) dedicated to the arts and humanities. As a research infrastructure, DARIAH's goal is to enable and empower researchers to conduct their research with the help of state-of-the-art tools and methods, in doing so in an internationally-connected way. This talk will examine how DARIAH creates a forum for knowledge exchange and collaboration for researchers, while also providing platforms such as the SSH Open Marketplace and DARIAH Campus which empower researchers to find the SSH tools and training materials that they need. It will also discuss working groups that give a forum for researchers across disciplines and countries to come together to confront common problems, such as Lexical Resources or Multilingual DH. It will also discuss our work with our sister infrastructure, CLARIN, in common initiatives such as the SSH Open Cluster. For instance, the SSH Open Marketplace, which was developed during the SSHOC project, was built in a Work Package led by DARIAH but with help from CLARIN colleagues, ensuring that at every step of the way it remains viable for linguistic uses. Together with CESSDA, CLARIN and DARIAH actively maintain the SSH Open Marketplace, ensuring that the data quality remains high and that the marketplace continues to adapt to meet the needs of researchers. DARIAH, as an international network, is also involved in other partnerships, such as a Cooperating Partnership with Princeton University which gave rise to a NEH-funded "New Languages for NLP: Bringing Linguistic Diversity in the Digital Humanities" workshop that is ongoing, with resulting training materials being uploaded into DARIAH-Campus so that everyone can benefit.

Christophe Parisse (University of Paris Nanterre, Modyco, Corli & Ortolang)
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.

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

In this talk we present mini-CIEP+, a sharable multilingual corpus of prose. mini-CIEP+ consists of the first part of ten different popular works of prose across many different languages (version 1.0 contains 35 languages, with more planned for the future). It is therefore almost entirely parallel, while a few subcorpora are comparable rather than parallel. Subcorpora typically contain 5750 sentences and almost 125K tokens. Subcorpora have dependency grammar annotation based on the Universal Dependencies standard (de Marneffe et al., 2021). It is shareable due to recent developments in German law, which allow researchers to share up to 15% of copy-righted material with a select group of people. We discuss these developments and detail the conditions under which mini-CIEP+ can be shared. We additionally describe future plans for further annotation and analysis of mini-CIEP+. These future plans centre around two major efforts: (1) annotating the corpus for information status using a newly built annotation scheme and (2) generating sentence alignment for the corpus, with the ultimate aim of annotation projection across subcorpora. The talk will highlight preliminary results for both projects.

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. Computational Linguistics, 47(2), 255–308. https://doi.org/10.1162/coli_a_00402

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

In this presentation, I will illustrate how cross-linguistic data formats (CLDF, Forkel et al. 2018) can be used to construct intuitively accessible online databases of different types of cross-linguistic data, and store them according to FAIR principles. After briefly introducing some of CLDF's design principles, I will focus on presenting the CINWA database for names of cultivated plants in South American Indigenous communities (Aguilar Panchi et al. 2023, Urban et al. 2023) as an example of a CLDF database. CINWA contains names for cultivated plants in Indigenous languages of South America, which we use in ongoing work on linguistic aspects of the transition from hunting and gathering to sophisticated forms of agriculture that occurred at different times in different parts of pre-Columbian South America. I will discuss some of the decisions we made in designing the database and extracting data, highlighting in particular the value of (i) consistency in the generation of characters with diacritics in the orthographies of Indigenous languages and (ii) full documentation of the original glosses in the lexicographic sources that we consulted, and then demonstrate the functionalities of CINWA available to users through the online interface available at http://www.cinwa.org/

Panchi, A., Michelle, E., Lee, S. & Brodetsky, E. (2022). CINWA – Database of Cultivated plants and their names in the indigenous languages of South America. www.cinwa.org.

Forkel, R., List, J-M., Greenhill, S. J., Rzymski, Ch., Bank, S., Cysouw, M., Hammarström, H., Haspelmath, M., Kaiping, G. A., & Gray, R. D. (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data 5*: 180205. https://doi.org/10.1038/sdata.2018.205

Urban, M., Michelle, E., Panchi, A., Lee, S. & Brodetsky, E. (2023). CINWA (database of terminology for cultivated plants in indigenous languages of northwestern South America): introducing a resource for research in ethnobiology, anthropology, historical linguistics, and interdisciplinary research on the neolithic transition in South America. Language Resources and Evaluation 57: 1787-1817. https://doi.org/10.1007/s10579-022-09628-x

Cyriel Mallart (LIDILE, Univ. de Rennes), Andrew Simpkin (NUI Galway, National Univ. of Ireland), Rémi Venant (LIUM, Université 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).

Linguistic interoperability within a unified architecture: the case of Analytics for Language Learning

Modern approaches to quantitative linguistics rely on large datasets. These datasets are representations of linguistic observations made up of features of various dimensions. Analyses rely on these data representations. As a result, the question of their construction is essential. In the case of quantitative research methods, datasets are built atomatically. This includes the sequential use of various types of NLP tools. These tools are the 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 paradgmitatic 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 vizualisation module of the architecture. Overall this process relies 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 aformentionned roles. These programmes are related to each other thanks to APIs and a data transfer queuing system. Automation and interoperability are the cornerstones of quantitative data processing.

Bird, S. & Loper, E. (2004). NLTK: The Natural Language Toolkit. The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics, 214–217.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. Behavior Research Methods, 48(4), 1227–1237.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Gaillat, T. (2022a). Language learning analytics: Designing and testing new functional complexity measures in L2 writings. Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning, 55–60.

Gaillat, T. (2022b). Language learning analytics: Designing and testing new functional complexity measures in L2 writings. 55. https://doi.org/10.3384/ecp190006

Kyle, K. (2016). Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication [Dissertation, Georgia State University].https://scholarworks.gsu.edu/alesl_diss/35

Li, J.-Y., Gaillat, T., & Richard, É. (2022). Exploring the use of dependency parsing in automatic erroneous collocation extraction in learner English. LCR2022, Padua, Italy. https://www.aclweb.org/anthology/2020.mwe-1.13

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages (arXiv:2003.07082). arXiv. https://doi.org/10.48550/arXiv.2003.07082

Rehm, G., Bontcheva, K., Choukri, K., Hajič, J., Piperidis, S., & Vasiļjevs, A. (Eds.). (2020). Proceedings of the 1st International Workshop on Language Technology Platforms. European Language Resources Association. https://www.aclweb.org/anthology/2020.iwltp-1.0

Sérasset, G., Witt, A., Heid, U., & Sasaki, F. (2009). Multilingual language resources and interoperability. Language Resources and Evaluation, 43(1), 1–14.

Straka, M., Hajič, J., & Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization,

Morphological Analysis, POS Tagging and Parsing. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 4290–4297. https://aclanthology.org/L16-1680

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

The Clarin infrastructure has been actively promoting the development of language resources, reference corpora and Natural Language Processing, for over a decade. As a result, academic and corporate researchers have now at their disposal a wide range of reference corpora for over 100 languages, from the Universal Dependencies corpora repository (available at https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-5150), together with state-of-the art NLP tools such as the UDPipe part-of-speech tagger/lemmatizer and functional dependency parser (Straka M. et al, 2016), available as a web application (https://lindat.mff.cuni.cz/services/udpipe/) or as a web service (https://lindat.mff.cuni.cz/services/udpipe/api/). The domain of Computer-Assisted Language Learning (CALL) has actively been experimenting with digital resources to foster -and optimize- language learning, as early as seminal works such as (Bitzer et al., 1961). In more recent approaches, such as (Aldabe et al., 2006; Borin & Saxena, 2005; Lee & Seneff, 2007; Mitkov et al., 2006; Perez-Beltrachini et al., 2012; Smith et al., 2010) and (Heck & Meurers, 2022) several projects have explored different aspects of using reference corpora and NLP tools, to automatically derive vocabulary questions and grammar exercises. In this presentation, we will address the topic of how to integrate reference corpora, language resources and NLP tools to teach linguistics. In other words, we will address computer-assisted metalinguistic competence acquisition, rather than language acquisition per se. We will present preliminary results of an ongoing research project aiming at deriving formative and evaluative self-correcting quizzes and other activities for syntax, by relying on existing reference corpora and NLP tools available from the Clarin infrastructure. We will present a software platform that enables teachers to generate Moodle compatible syntax quizzes and other activities, by leveraging annotations present in manually-revised reference corpora, as well as from NLP tools such as UDPipe. Our platform aims both at reducing manual edition to a minimum and at overcoming the subjectivity (and errors) associated with manually created exercises. Here, we report experiments conducted on French, Spanish and English, although the approach can be extended to other languages, since CONLL-U corpora are readily available from universaldependencies.org for over 100 different languages.

TUTORIAL SESSION (registration required*)

Christophe Parisse, Alina Tsikulina, Adélie Buclier & Eva Soroli
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.

*Registration: This last tutorial session on basic tools for discourse analysis is limited to a maximum of 15 participants. To register, please send an email to efstathia.soroli@univlille.fr by providing 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). Accepted participants will be notified by e-mail and will be asked to download the CLAN software to their machines before coming to this tutorial session with their laptop. Please follow the instructions here: https://dali.talkbank.org/clan/.
Registrations for this tutorial will be open until: August 16, 2024. Notifications will be sent by the end of August.