Title	CLARIN Annual Member and Observer Reports 2023
Version	1
Author(s)	KB, NCF
Date	6 August 2024
Status	For distribution
Distribution	NCF, GA, BoD
ID	CE-2024-2441

CLARIN Member and Observer Reports 2023

Introduction

This document provides an overview of the organisation and activities for each CLARIN ERIC member and observer. It lists the coordinator, consortium, core user communities, website and key achievements, **for the period November 2022 to November 2023**. During this period, Spain joined as full member (September 2023), and Switzerland as observer (January 2023). This brings the total number of members to 23, and the number of observers to 3.

On the following pages, the reports are listed by country alphabetically.

1. Austria
2. Belgium
3. Bulgaria
4. Croatia
5. Cyprus
6. Czech Republic
7. Denmark
8. Estonia
9. Finland
10. Germany
11. Greece
12. Hungary
13. Iceland
14. Italy
15. Latvia
16. Lithuania
17. The Netherlands
18. Norway
19. Poland
20. Portugal
21. Slovenia
22. South Africa (Observer)
23. Spain
24. Sweden
25. Switzerland (Observer)
26. United Kingdom (Observer)

Austria/CLARIAH-AT	CLARIAH-AT

Coordinator name and institution

Tanja Wissik (Austrian Academy of Sciences, ACDH-CH)
Walter Scholger (University of Graz, ZIM-ACDH, Graz), deputy

List of consortium partners (with centre types hosted)

1. Österreichische Akademie der Wissenschaften (ARCHE – CLARIN B-centre; Phonogrammarchiv – CLARIN K-centre)
2. Österreichische Nationalbibliothek (ÖNB)
3. Universität für Weiterbildung Krems
4. Universität Graz (GAMS – CLARIN B-centre)
5. Universität Innsbruck
6. Universität Klagenfurt
7. Universität Salzburg
8. Universität Wien (TRTC – CLARIN K-centre)

Core user communities

Digital humanities scholars, Linguists and language scholars, Historians, Translation scholars, Literary scholars, Librarians, Archivists

Website

https://clariah.at
https://twitter.com/AtClariah

Key achievements

Events
a. *DH 2023* (2023, July 10-14) - Graz, Austria. https://dh2023.adho.org/
b. *'Computer Vision for Digital Humanists' Winter School* (2023, February 8-10) - https://static.unigraz.at/fileadmin/_files/_gewi_sites/_informationsmodellierung/PDF/uni-graz-informationsmodellierung-UPDATED_2023_CWWinterSchool-Programme.pdf
c. *Österreichisches Treffen zur Digitalen Linguistik* (2023, December 8) - Österreichische Linguistik-Tagung (ÖLT) 2023. https://clariah.at/de/digitale-linguistik-rezente-entwicklungen-in-osterreich-olt-2023
d. *Österreichisches Treffen zu Sentimentinferenz (ÖTSI)* (2023, December 8) - Österreichische Linguistik-Tagung (ÖLT) 2023. https://evsl.univie.ac.at/digital-philology/oesterreichisches-treffen-zu-sentimentinferenz-oetsi/

e. *DigEdTnT: Meet the Tools and Developers* [Workshop] (2023, February 23-24) - https://informationsmodellierung.uni-graz.at/de/veranstaltungsuebersicht/vergangene-veranstaltungen/

Projects

a. *Digital Edition Creation Pipelines: Tools and Transitions (DigEdTnT)* - https://digedtnt.github.io/
b. *LadderWeb* - https://ladder3ca-test.literaturtirol.at/#/home

Resources (datasets, tools, services)

a. Erjavec, T. et al. (2023). *Multilingual comparable corpora of parliamentary debates ParlaMint 4.0*. Slovenian language resource repository CLARIN.SI http://hdl.handle.net/11356/1859 [ParlaMint-AT: Austrian dataset as part of ParlaMint]

Publications

a. Baumann, A., Hofmann, K., Marakasova, A., Neidhardt, J., & Wissik, T. (2023). Semantic micro-dynamics as a reflex of occurrence frequency: A semantic networks approach. *Cognitive Linguistics, 34*(3-4), 533-568. https://doi.org/10.1515/cog-2022-0008
b. Nicolaou, A., Luger, D., Decker, F., Renet, N., Christlein, V., & Vogeler, G. (2023). Efficient annotation of medieval charters. In M. Coustaty, A. Fornés (Eds.), *Document analysis and recognition – ICDAR 2023 workshops*. Springer. https://doi.org/10.1007/978-3-031-41498-5_20
c. Hofeneder, P. (2023). The circulation of knowledge vs the mobility of translation or how mobile are translators and translations?. In I. Feinauer, A. Marais, & M. Swart (Eds.), *Translation flows: Exploring networks of people, processes and products* (pp. 23-42). John Benjamins. https://doi.org/10.1075/btl.163.02hof
d. Scholger, W., & Gietz, P. (2023). Complexities in intellectual property and the application of regulatory frameworks in specific research scenarios. In E. Tóth-Czifra, M. Błaszczyńska, & F. Gelati (Eds.), *Research data management for arts and humanities: Integrating voices of the community* (pp. 56-75). DARIAH-EU.
e. Wissik, T., & Hofman, K. (2023). Emotion and Abstractness in Austrian Parliamentary Discourse. In K. Lindén, J.Niemi, & T. Kontino (Eds.), *Proceedings of the CLARIN Annual Conference 2023*, 159-163.

Belgium/CLARIN BE

Coordinator name and institution

Vincent Vandeghinste, Instituut voor de Nederlandse Taal (Dutch Language Institute) and KU Leuven

List of consortium partners (with centre types hosted)

1. Artificial Intelligence Laboratory at Vrije Universiteit Brussel
2. Centre for Computational Linguistics and Psycholinguistics (CLIPS) at University of Antwerp
3. Centre for Computational Linguistics (CCL) at KU Leuven
4. Human Computer Interaction (HCI) at KU Leuven
5. Instituut voor de Nederlandse Taal (B-centre + K-centre)
6. Internet Technology and Data Science (imec-IDLab) at Ghent University
7. Itec Educational Technology at KU Leuven (Kortrijk)
8. Language Intelligence and Information Retrieval lab (LIIR) at KU Leuven
9. Language and Translation Technology Team (LT3) at Ghent University
10. Namur Digital Institute at Université de Namur
11. Pole Linguistique (PLIN) at Université Catholique de Louvain (K-Centre)
12. Processing Speech and Images (ESAT-PSI) at KU Leuven

Core user communities

Linguistics and language scholars, Digital humanities scholars

Website

https://clarin-be.ivdnt.org/
https://clariahvl.hypotheses.org/

Key achievements

Events
   a. *DH Benelux 2023* (2023, May 13 - June 2) - https://lab.kb.nl/agenda/dh-benelux-2023
   b. *CLARIN Annual Conference 2023* [host] (2023, October 16-18) - https://www.clarin.eu/event/2023/clarin-annual-conference-2023

c. *Learner Corpus Research Summer School,* 5th edition (2023, July 3-7) - https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpus-research-summer-school-2023.html

Projects

a. *SignON* - Sign Language translation project, which resulted in several Sign Language datasets. https://signon-project.eu/
b. *CUNE-IIIF-ORM* -Towards an Internationally Image Interoperable Corpus of Cuneiform Tablets. https://www.kmkg-mrah.be/en/scientific-research/cune-iiif-orm

Resources (datasets, tools, services)

a. Several Sign Language Datasets
b. *MATEO: MAchine Translation Evaluation Online*. https://mateo.ivdnt.org/
c. *LLM for Ancient and Byzantyne Greek*, fine-tuned for part-of-speech tagging. Code/model: https://huggingface.co/colinswaelens/DBBErt [not yet in CLARIN catalogue]
d. *Dutch event detection in newspaper text*. Data and code: https://github.com/NewsDNA-LT3/.github [not yet in CLARIN catalogue]
e. New resources have been made available via https://corpora.uclouvain.be/catalog/ and importantly, for us, this catalogue is now harvested by the VLO.

Publications

a. De Sisto, M., Shterionov, D., Soetemans, L., Vandeghinste, V., & Brosens, C. (2023). NGT-HoReCo and GoSt-ParC-Sign: Two new Sign Language: Spoken Language parallel corpora. In K. Lindén, J.Niemi, & T. Kontino (Eds.), *Proceedings of the CLARIN Annual Conference 2023*. https://www.clarin.eu/sites/default/files/CLARIN2023_3.1.2_De_Sisto.pdf
b. Vandeghinste, V.,Mathijssen J., Peters, E., & Wambacq, P. (2023). A Spoken Academic Belgian Dutch Corpus. In K. Lindén, J.Niemi, & T. Kontino (Eds.), *Proceedings of the CLARIN Annual Conference 2023*.
c. Vanroy, B., Tezcan, A., & Macken, L. (2023). MATEO: MAchine Translation Evaluation Online. In M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, H. Moniz, et al. (Eds.), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation* (pp. 499–500). European Association for Machine Translation (EAMT).
d. Swaelens, C., De Vos, I., & Lefever, E. (2023). DBBErt: Part-of-Speech Tagging of Pre-Modern Greek Text. In K. Lindén, J.Niemi, & T. Kontino (Eds.), *Proceedings of the CLARIN Annual Conference 2023* (pp. 155–58).
e. Paquot, M., König, A., Stemle, E. & Frey, J.C. (2023). *Core Metadata Schema for Learner Corpora* (Version 1) [Data set]. Open Data @ UCLouvain. https://doi.org/10.14428/DVN/4CDX3P

Training materials

a. Taal en AI: 5 videos about Language and AI https://www.youtube.com/playlist?list=PLpQLxGxT7ucH_hGuy9Kozu-AdW_yx3ZYO
b. Annotation guidelines Sentiment Analysis:

- Singh, P., Maladry, A., & Lefever, E. (2023). Annotation guidelines for labeling emotion in multilingual tweets. Ghent University. Faculty of Arts, Humanities and Law. LT3.
- De Geyndt, E., De Clercq, O., Van Hee, C., Lefever, E., & Hoste, V. (2023). Annotation guidelines for labeling aspect-based sentiment analysis and aspect-based emotion analysis in the SentEMO project: version 1.0. Ghent University. Faculty of Arts, Humanities and Law. LT3.

Bulgaria/CLaDA-Bulgaria

Coordinator name and institution

Kiril Simov (Institute of Information and Communication Technologies, Bulgarian Academy of Sciences)

Petya Osenova (Institute of Information and Communication Technologies, Bulgarian Academy of Sciences), deputy

List of consortium partners (with centre types hosted)

1. Institute of Information and Communication Technologies at the Bulgarian Academy of Sciences
2. Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences
3. Ontotext Corp
4. Sofia University 'St. Kliment Ohridski'
5. New Bulgarian University
6. Konstantin Preslavsky University – Shumen
7. Bulgariana – NGO promoting CH in Bulgaria
8. South-West University 'Neofit Rilski' – Blagoevgrad
9. Sirma Media
10. Cyrillo-Methodian Research Centre at the Bulgarian Academy of Sciences
11. Institute of Balkan Studies and Centre of Thracology 'Alexander Fol' at the Bulgarian Academy of Sciences
12. Institute of Ethnology and Folklore Studies with Ethnographic Museum at the Bulgarian Academy of Sciences
13. Burgas Free University
14. 'Ivan Vazov' Public Library – Plovdiv
15. Sofia History Museum

Core user communities

Linguists and language scholars, Digital humanities scholar, Sociologists, Language teaching scholars, Language teachers, NLP scholars, Librarians

Website

https://clada-bg.eu/

Key achievements

Events

a. *Language Technologies and Digital Humanities: Resources and Applications (LT⊠DH-RA)* (2023, May 10-12) - CLaDA-BG 2023 Conference, Sofia, Bulgaria https://clada-bg.eu/en/dissemination/events/international-clada-bg-conference-2023.html
b. CLARIN Annual Conference 2023 (2023, October 16 -18) - https://www.clarin.eu/event/2023/clarin-annual-conference-2023
c. Iliev, D., & Kalfina, V. (2023, September 19-23). *Encoding Cicero: The Latin-Bulgarian electronic dictionary: challenges and perspectives* [Presentation]. Perceptions of Caucasus Pre-Conference Workshop, Tbilisi State University.
d. *The Digital Collections of Greek and Latin Inscriptions from Bulgaria* [Poster] (2023, June 19-23). 10th Joint Meeting ECFN-Nomisma.org, Sofia, Bulgaria.
e. Sharankov, N. & Iliev, D. (2023, April 24-26). *The digital collections of Greek and Latin inscriptions from Bulgaria: Challenges and perspectives* [Presentation]. ENCODE/Epigraphy.info VII Workshop, Leuven.

Projects

a. *Historical Hameleons* (within CLaDA-BG) - https://historical-chameleons.clada-bg.com/en/home/
b. *BERON: The official online spelling dictionary of Bulgarian* (collaboration with teams outside CLaDA-BG) - https://beron.mon.bg/beron

Resources (datasets, tools, services)

a. *Event Corpus* of Bulgarian
b. *BERT model* for Bulgarian
c. *GPT-2 model* for Bulgarian
d. *CultIS* - a web-based software platform for intelligent digital management and presentation of large datasets and knowledge in the fields of culture, humanities and social sciences, https://cultis.math.bas.bg/en
e. *Digital Library of Plovdiv National Library* - https://digital.libplovdiv.com/

Collaborations (notable collaborations with other academic institutions, and/or GLAM sector or industry)

Associate partners:
Bulgarian Encyclopaedia Scientific Information Centre, BAS
Prof. Marin Drinov Publishing House of BAS
The Central Library of the Bulgarian Academy of Sciences
Museum of Folk Art Crafts and Applied Arts - Troyan

Publications

a. Angelov, Z., Simov, K., Osenova, P., & Kancheva, Z. (2023). The CLaDA-BG Dictionary Creation System: Specifics and Perspectives. *Selected papers from the CLARIN Annual Conference 2022* (pp. 12-22). CLARIN. https://doi.org/10.3384/ecp198002

b. Marinova, I., Simov, K., & Osenova, P. (2023). Transformer-Based Language Models for Bulgarian. *Proceedings of Recent Advances in Natural Language Processing, Varna, Septemper 4–6, 2023* (pp. 712-720). NCOMA Ltd. https://doi.org/10.26615/978-954-452-092-2_077

c. Hardalov, M., Atanasova, P., Mihaylov, T., Angelova, G., Simov, K., Osenova, P., Stoyanov, V., Koychev, I., Nakov, P., & Radev, D. (2023). bgGLUE: A Bulgarian General Language Understanding Evaluation Benchmark. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Vol. 1*(Long Papers), 8733-8759. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.487

d. Raykovska M., Jones K., Klecherova H., Alexandrov S., Petkov N., Hristova T., & Ivanov G. (2023). Implementing Digital Documentation Techniques for Archaeological Artifacts to Develop a Virtual Exhibition: the Necropolis of Baley Collection. Under review for *JCAA PROCEEDINGS of the International CLaDA-BG Conference 2021 Language Technologies and Digital Humanities in Bulgaria: 6-7 Sept. 2021, Varna*. Incoma Ltd. https://clada-bg.eu/images/Proceedings-CLaDA-BG-2021.pdf

e. Iliev, D., & Boeva, E. (2023). From stone to screen: The Telamon database of ancient inscriptions in Greek from Bulgaria. *AIP Conference Proceedings, 2939*(1). https://doi.org/10.1063/5.0178753

Croatia/HR-CLARIN

Coordinator name and institution

Marko Tadić, University of Zagreb, Faculty of Humanities and Social Sciences

List of consortium partners (with centre types hosted)

1. University of Zagreb, Faculty of Humanities and Social Sciences (FFZG)
2. University of Zagreb Computing Centre (SRCE)
3. University of Zagreb, Faculty of Electrical Engineering and Computing (FER)
4. Institute of Croatian Language and Linguistics (IHJJ)
5. Old-Church Slavonic Institute (STIN)
6. National and University Library (NSK)
7. Croatian Academy of Sciences and Arts, Library (HAZU)
8. Croatian Language Technologies Society (HDJT)

Core user communities

Linguistics and language scholars, Digital humanities scholars, Information and communication scientists

Website

https://www.clarin.hr/

Key achievements

Events
a. *EACL2023* (2023, May 2-6) - [Co-organisation], Dubrovnik.
b. *CLARC2023* (2023, September 28-30) - [Co-organisation], Rijeka.
c. *DeriMo2023* (2023, October 5-6). - [Co-organisation], Dubrovnik.
d. *Croatian old literature: State of the art, contemporary access to language materials and their digitalisation* (2023, December 14) - [Colloquium/presentation], Zagreb.

Projects

a. *CLEOPATRA, MSC ITN* - on cross-lingual event-centric open analytics. [Thorsten Trippel (CLARIN ERIC representative) gave a keynote speech at the final CLEOPATRA workshop in Hannover, May 2023.]
b. *NLTP* - CEF project on building a national language technology platform.
c. *European Language Equality 2* - project of the European Parliament.
d. *European LT Domains 2023 (EuLTDom2023)* - European Language Equality 2 cascaded project.
e. *Computing facilities for LT (ComFac4LT)* - European Language Equality 2 cascaded project.

Resources (datasets, tools, services)

a. *CroSentiNews 2.0* - A Sentence-Level News Sentiment Corpus.
b. *Croatian Film Review Dataset (Cro-FiReDa)* - A Sentiment Annotated Dataset of Film Reviews.

Collaborations (notable collaborations with other academic institutions, and/or GLAM sector or industry)

Collaboration with the Central State Office for the Development of Digital Society of the Republic of Croatia regarding the participation of Croatia in ALT-EDIC.

Publications

a. Alves, D., Bekavac, B., Zeman, D., Tadić, M. (2023). Corpus-based synta*ctic* typological methods for dependency parsing improvement. In E. Vylomova (Ed.), *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, (pp. 76-88.). SIGTYP and Association for Computational Linguistics.
b. Tadić, M. (2023). Language Report Croatian. In G. Rehm, & A. Way (Eds.), *European Language Equality: Cognitive Technologies*. Springer. https://doi.org/10.1007/978-3-031-28819-7_9
c. Tadić, M. (2023). Can large language models tell us something about derivation processes?. *Proceedings of the Fourth International Workshop on Resources and Tools for Derivational Morphology (DeriMo2023)* (pp. 33-38). Croatian Language Technologies Society.
d. Thakkar, G., Tadić, M., & Rigau, G. (2023). ComFac4LT – Computing facilities for language technologies. *Central European Conference on Information and Intelligent Systems: Proceedings* (pp. 325-340). University of Zagreb, Faculty of Organisation and Informatics. 2023.
e. Vasiļevskis, A., Ziediņš, J., Tadić, M., Motika, Ž., Fishel, M., Barkarson, B., Borg, C., Aquilina, K., & Spiteri, D. (2023). National Language Technology Platform (NLTP): The Final Stage. In G. Corpas Pastor, R. Mitkov, et al. (Eds.), Proceedings of the International Conference HiT-IT 2023 (pp. 203-208). INCOMA Ltd. https://doi.org/10.26615/issn.2683-0078.2023_019

Cyprus/CLARIN-CY

Coordinator name and institution

Marinos Ioannides, UNESCO Chair on Digital Cultural Heritage, Cyprus University of Technology

List of consortium partners (with centre types hosted)

1. Cyprus Digital Library/ Ministry of Education, Sport and Youth
2. University of Cyprus
3. University of Nicosia
4. Bank of Cyprus, Culture Association
5. Department of Antiquities

Core user communities

Digital humanities scholars, Cultural information and communication scientists

Website

https://digitalheritagelab.eu/
https://erachair-dch.eu/

Key achievements

Cyprus has actively contributed to the CLARIN infrastructure by promoting digital cultural heritage initiatives and engaging with key user communities. Notable achievements include participation in various projects, organisation of impactful events, and production of significant publications. The efforts have enhanced the visibility and integration of CLARIN within the cultural heritage sector, fostering collaboration and innovation in digital humanities.

Events

a. *European Researchers' Night: Theme Mission Connect* (2023, September 29) - Nicosia State Fair https://erachair-dch.eu/event/in-person-symmetochi-sti-vradia-toy-ereyniti-2023/
   - **Mission:** The European Researchers' Night is an annual event aimed at educating the public about the world of science and research while strengthening the public image of researchers. The event is organized under the EU-funded Project CONNECT (Grant Agreement No. 101060913) and has been hosted by the Research and Innovation Foundation in Cyprus since 2006, in collaboration with key actors in the local research and innovation ecosystem, including public and private universities, research centers, non-governmental organizations, and businesses.

That year's theme, Climate-Neutral and Smart Cities, aimed to nurture a culture of research and innovation, raising awareness of the research activities in Cyprus. Scientists, academics, researchers, and entrepreneurs presented their work to the public, engaging visitors of all ages with interactive experiments, games, discussions, and awards.

- **Impact:** The UNESCO and ERA Chairs on Digital Cultural Heritage based at the Cyprus University of Technology (CUT) participated actively. They showcased the holistic digital documentation of cultural heritage, including 3D digitization projects like the traditional village of Fikardou. Visitors experienced virtual tours of Fikardou, digital explorations of local gastronomy, and the use of the Antikythera Mechanism application .

b. *Final Event Meta-Pandemic GLAMers in Cooperation with the EU Project MNEMOSYNE (H2020 ERA CHAIR)* (2023, February 13) - Limassol, Cyprus https://erachair-dch.eu/event/conference-final-event-meta-pandemic-glamers/

- **Mission:** The GLAMers project, coordinated by CUT through the UNESCO Chair in Digital Cultural Heritage, aimed to support galleries, libraries, archives, and museums (GLAMs) in their digital transformation, particularly in response to the COVID-19 crisis. The project involved youth in enhancing GLAM activities and evaluated the impact on social changes in the relationship between GLAMs and youth.
- **Impact:** The event brought together cultural and creative institutions, youth communities, and digital transformation experts. It featured speeches by notable figures such as the Deputy Minister of Culture of Cyprus and the Rector of CUT. The event provided a platform for sharing project outcomes and exploring innovative approaches to strengthen GLAM-youth relationships in a post-pandemic world.

c. *The 16th International Conference for Cultural Tourism in Europe* (2023, October 18-21) - Pafos, Cyprus, https://erachair-dch.eu/event/in-person-the-16th-international-conference-for-cultural-tourism-in-europe/

- **Mission:** The UNESCO Chair on Digital Cultural Heritage at CUT, through the Mnemosyne project, participated in this conference focused on smart tourism, cultural heritage, and sustainability. Organized by the European Cultural Tourism Network (ECTN), the conference aimed to address challenges and opportunities in smart, sustainable, and responsible tourism development.
- **Impact:** Mnemosyne research fellows presented on topics such as the holistic digital documentation of Saint Neophytos Monastery and the use of smart technologies in cultural heritage tourism. The conference facilitated the exchange of best practices and provided practical recommendations for promoting cultural heritage through innovative digital methods.

d. *Project MNEMOSYNE Final Event (H2020 ERA CHAIR)* (2023, December 7-8) - Limassol, Cyprus https://erachair-dch.eu/event/mnemosyne-project-announces-its-final-event/

- **Mission:** The final event of the Horizon ERA Chair MNEMOSYNE project highlighted the project's achievements in holistic digital documentation of cultural heritage over five years. The event included presentations, round table discussions, and demonstrations of practical applications of the project's innovations.
- **Impact:** The event was attended by a diverse group of experts and stakeholders, including government officials, legal experts, and academic professionals.

Presentations covered various aspects of the Mnemosyne project's work, showcasing its impact on digital cultural heritage practices and policies

e.   *CLARIN Infrastructure Promotion*
   -   Throughout these events, the CLARIN infrastructure was presented and promoted to inform the audience about its capabilities and benefits. This effort was part of the overarching strategy to integrate CLARIN into broader cultural heritage and research initiatives, thereby enhancing its visibility and adoption within the research community.
   -   This report highlights the successful organization and impact of these key events, demonstrating the active role of the Cyprus University of Technology and its associated projects in promoting digital cultural heritage and innovative research practices.

Projects

a.   *EU H2020 ERA Chair on Digital Cultural Heritage MNEMOSYNE* (2019 - 2023) - https://erachair-dch.eu/
   -   Mnemosyne is a project for a single-stage Coordination and Support Action submitted under WIDESPREAD-03-2017-ERA Chairs. Context Cultural Heritage is a strategic resource for Europe with high cultural, social, environmental and economic value. The era of Digital Cultural Heritage (DCH) is now well underway and the European research resource for DCH has grown significantly in recent years. Cultural Heritage has a strategic resource for growth, job creation and a role to play for more cohesive and sustainable regions in the EU.
b.   *ERASMUS+ The GLAMers* (2021-2023) - https://glamers.eu/
   -   The main goal of the study is to produce a framework that would enable cultural heritage professionals and institutions and other custodians of cultural heritage, providers of 3D digitisation services for cultural heritage and other researchers from the area of 3D digitisation technologies for cultural heritage to define and produce high quality in the context of specific 3D digitisation projects for tangible cultural heritage.
c.   *ANCHISE* (2023 - 2026) - https://www.anchise.eu/
   -   ANCHISE – 'Applying New solutions for Cultural Heritage protection by Innovative, Scientific, social and economic Engagement' aims at offering to European societies efficient methods, knowledge, and toolkit to enhance the protection of cultural heritage against looting and illicit trafficking.
d.   *EUreka3D* (2023 - 2025) - https://eureka3d.eu/
   -   EUreka3D will prove innovative approaches to digital transformation of CH sector, based on the provision of services and tools, compliant with Europeana CSP, and powered by advanced computing and storage resources. EOSC registered services for authentication and authentication, guidelines and recommendations for high-quality 3D digitization, representation of holistic documentation in advanced metadata information, all this will be assessed in a pilot where 4 CHIs participate with their contents.
e.   *ERASMUS+ ARTEST* (2023-2025) - https://artest-project.eu/

- The project aims at reinforcing education in the field of the Humanities by adopting digital methods of research and education and harmonizing the development of both soft and digital skills during the study process.

Collaborations (notable collaborations with other academic institutions, and/or GLAM sector or industry)

a. UNESCO Chair in History and Heritage at the State of the University of Haiti
b. National Museums of World Culture
c. Heritage Malta
d. E-Archiving Initiative
e. Korea National University of Cultural Heritage - KNUCH

Publications

a. Argyridou, E., Karaoli, A., Hadjiathanasiou, M., Karittevli, E., Panagi, I., Mateou, M., Ioannides, M., Patias, P., & Efstathiou, K. (2023). The first attempt for standardisation in 3D digitisation: The EU study on quality in 3D digitisation of tangible cultural heritage. *CIPA 2023 Conference Proceedings*, 103–109. https://doi.org/10.5194/isprs-archives-XLVIII-M-2-2023-103-2023
b. Argyridou, E., Efstathiou, K., Hadjiathanasiou, M., Ioannides, M., Karaoli, A., Karittevli, E., Mateou, M., Panagi, I., & Samara, N. (2023). Digital holistic documentation of cultural heritage: Challenges and risks, towards shaping the future. *CIPA 2023 Conference Proceedings*, 95-102. https://doi.org/10.5194/isprs-archives-XLVIII-M-2-2023-95-2023
c. Argyridou, E., Rizopoulos, O., Hadjiathanasiou, M., Ioannides, M., Samara, P., Karittevli, E., Panayi, I., Mateou, M., Koulafeti, I., Koundouris, M., Efstathiou, K., Baker, D., & Savva, G. (2023). *Conference Proceedings: International Conference on Cultural Tourism Advances (ICCTA)*. Springer Nature.
d. Argyridou, E., Samara, P., Ioannides, M., Hadjathanasiou, M., Karittevli, E., Koulafeti, I., Panayi, I., & Efstathiou, K. (2023). Research Methodologies for Digital Holistic Documentation of Cultural Heritage. *Conference Proceedings EVA London 2023*.
e. Karittevli, E., Kokkinos, N., Ioannides, M., & Efstathiou, K. (2023). Conference Proceedings EVA Berlin 2023.

Czech Republic/
LINDAT/CLARIAH-CZ, CNC

Coordinator name and institution

Eva Hajičová, Charles University (CU)

List of consortium partners (with centre types hosted)

1. Charles University (CLARIN Centers B, K)
2. Masaryk University
3. University of West Bohemia
4. Czech Language Institute of the Czech Academy of Sciences
5. Institute of History of the Czech Academy of Sciences
6. Institute of Philosophy of the Czech Academy of Sciences
7. Library of the Czech Academy of Sciences of the Czech Academy of Sciences
8. Moravian Library in Brno
9. National Library of the Czech Republic
10. National Film Archive
11. National Gallery Prague
12. Masaryk Institute and Archives of the Czech Academy of Sciences
13. Terezín Initiative Institute
14. Terezín Memorial
15. The National Archives

Czech National Corpus (CNC) - funded separately (CLARIN Centre E)
1. Charles University, Faculty of Arts

Core user communities

Digital humanities scholars, Computational linguists, NLP scholars, Corpus linguists, historians, sociologists, ethnologists, librarians, holocaust researchers

Website

https://lindat.cz
https://korpus.cz

Key achievements

Events
a. *MALACH Center for Visual History, 13th Anniversary Conference* (2023, January 30) - Prague, Czech Republic
b. *SIGDIAL-INLG* (2023, September 11-15) - Prague, Czech Republic
c. *LDS workshop on LLMs (O. Bojar),* (2023, June 6) - online
d. *CNC user workshop (*2023, June 6) - Using CNC in teaching Czech as a foreign language

e. *CNC user workshop* (2023, November 25) - Teaching of Czech for secondary schools

Projects

a. *CLS INFRA project (Computer Literary Studies Infrastructure, LINDAT+CNC)* - EC Project No. 101004984 (Institute of Polish Language at the Polish Academy of Sciences, Poland; University of Potsdam, Germany; Austrian Academy of Sciences, Austria; National University of Distance Education, Spain; École Normale Supérieure de Lyon, France; Humboldt University of Berlin, German; DARIAH ERIC, France; Ghent Centre for Digital Humanities, Ghent University, Belgium; Belgrade Centre for Digital Humanities, Serbia; Huygens Institute for the History of the Netherlands (Royal Netherlands Academy of Arts and Sciences), Netherlands; Trier Center for Digital Humanities, Trier University, Germany; Moore Institute, National University of Ireland Galway, Ireland; The Trinity Centre for Digital Humanities, Trinity College Dublin, Ireland)
b. *COST Action UniDive* – EC Project No. CA21167 (cooperation with 250 researchers in 36 COST countries + 5 more countries).
c. *HPLT High-performance Language Technology (2022-2025)* - EC project No. 101070350, coordinated by CU (Univ. of Edinburgh, Univ. of Turku, Univ. of Helsinki, University of Oslo, CESNET CZ, Sigma2 Norway, PrompsIt Spain)
d. *ParlaMint* - continued project participation (led by CLARIN-SI)
e. *European Language Equality 2* - EC Project No. 101075356 (Dublin City University (DCU) – Coordinator; Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI); Universidad Del Pais Vasco/Euskal Herriko Unibertsitatea/University of the Basque Country (UPV/EHU); Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis – Athena Research and Innovation Center in Information, Communication and Knowledge Technologies, Institute for Language and Speech Processing (ISLP); European Federation of National Institutes for Language (EFNIL); Réseau européen pour l'égalité des langues/European Language Equality Network (ELEN)
f. *ELRC (CEF project),* (2015-2023) -EC Project No. CEF SMART 2019/1083, Ref.: LC-01325001; 30 partners from EU, European Language Resource Consortium
g. *LAPPS Grid – CLARIN cooperation* - (Univ. of Tübingen, Brandeis University (MA, USA), Vassar College (NY, USA)), Mellon Foundation

Resources (datasets, tools, services)

a. Zeman, D., et al. (2023, November). *Universal Dependencies 2.13* [Data/software]. LINDAT/CLARIAH-CZ digital library, Institute of Formal and Applied Linguistics, Charles University in Prague. http://hdl.handle.net/11234/1-5287
b. Novák, M., Popel, M., Žabokrtský, Z., Zeman, D., Nedoluzhko, A., et al. (2023, February). *CorefUD 1.1* [Data/software]. Univerzita Karlova. http://hdl.handle.net/11234/1-5053
c. Urešová, Z., et al. (2023). *SynSemClass 5.0.* LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-5230
d. New version of the Mapka application for working with spoken and dialectal corpora. It fea-tures en-hanced de-sign, more in-tu-itive in-ter-face, sig-nif-i-cant-ly enriched con-tent and editing mode for administration. (released 6. 9. 2023)

e. Complete overhaul of the KWords application for identification of keywords that includes both its frontend and backend. It employs UDPipe for annotation, and thus supports approximately 20 languages. (released 14. 11. 2023)

Collaborations (notable collaborations with other academic institutions, and/or GLAM sector or industry)

a. *CELL* - Contextual Machine Learning of Language Translations, funded by CELSA, project No. CELSA/19/018 (Univ. of Leuven)
b. *NG-NLG* - Next-Generation Natural Language Generation, ERC EC project No. 101039303. EC project but no formal cooperation (informally, Heriot-Watt Univ., Edinburgh, Scotland, UK)
c. *MEMORISE* - Virtualisation and Multimodal Exploration of Heritage on Nazi Persecution, EC Project No. 101061016 (Syddansk Universitet, Denmark; Stichting Herinneringscentrum Kamp Westerbork, Netherlands; Fluxguide Ausstellungssysteme Gmbh, Austria; Semantic Web Company Gmbh, Austria; The Hebrew University Of Jerusalem, Israel; Eodyne Systems, Spain; Koninklijke Nederlandse Akademie Van Wetenschappen – Knaw, Netherlands; Stichting Radboud Universiteit Netherlands; Stiftung Niedersächsische Gedenkstätten, Germany)
d. *RESQ Plus* - Comprehensive solutions of healthcare improvement based on the global Registry of Stroke Care Quality, EC Project No. 101057603 coordinated by UZIS ČR/Charles Univ. (Aalborg Universitet, Denmark; Alana Ai Sa Limited; Boehringer Ingelheim International GmbH, Germany; Chino Srl, Italy; Ethniko Kai Kapodistriako Panepistimio Athinon, Greece; Fakultni Nemocnice U Sv. Anny V Brně, Czechia; Fundacio Hospital Universitari Vall D'hebron - Institut De Recerca, Spain; Health Management Institute; Instytut Psychiatrii I Neurologii, Poland; Masarykova Univerzita, Czechia; Nogoprofilna Bolnitsa Za Aktivno Lechenie Po Nevrologia I Psihiatria Sv Naum, Bulgaria; Ontotext, Bulgaria; Spitalul Universitar De Urgenta Bucuresti, Romania; Stroke Alliance For Europe, Belgium; Technological University Dublin, Irerland; Timelex, Belgium; Universidad De Murcia, Spain; World Stroke Organization, Switzerland)
e. *Building a Broad Infrastructure for Uniform Meaning Representations* - NSF CCRI-CISE, Award #2213804 (Brandeis University, University of New Mexico, University of Colorado Boulder); also supported in part by the Czech Ministry of Education, Project No. LUAUS23283 'UMR'.

Publications

a. Erjavec, T., Ogrodniczuk, M., Osenova, P. et al. (2023). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation 57*(1), 415–448. https://doi.org/10.1007/s10579-021-09574-0
b. Chromá, A., Sláma, J., Matiasovitsová, K., & Treichelová, J. (2023). A morphologically annotated longitudinal corpus of spoken Czech child–adult interactions. *Language Resources and Evaluation 57*(1) https://doi.org/10.1007/s10579-023-09710-y
c. Mikulová, M. (2023). Expressing measure in Czech (Corpus-based study). *Jazykovedný časopis / Journal of Linguistics 74*(1), 108-118. SAP: Slovak Academic Press.
d. Jelínek, T. (2023). Morphological tagging and lemmatization of spoken corpora of Czech. In: K. Ekštein, F. Pártl, M. Konopík (Eds.), *Text, Speech, and Dialogue. TSD 2023. Lecture*

Notes in Computer Science 14102, 154-163. Springer. https://doi.org/10.1007/978-3-031-40498-6_14

e. Rosen, A. (2023), The InterCorp parallel corpus with a uniform annotation for all languages. *Jazykovedný časopis 74*(1), 254–265. 10.2478/jazcas-2023-0043

Denmark/ CLARIN-DK

Coordinator name and institution

Costanza Navarretta (University of Copenhagen)
Bente Maegaard (University of Copenhagen), deputy

List of consortium partners (with centre types hosted)

1. University of Copenhagen, CLARIN B-centre
2. Danish Language Council, and Department of Nordic Studies and Linguistics, University of Copenhagen: K-centre DANSK (Danish Help desk)

Core user communities

Digital humanities scholars, Linguistics and language scholars, literacy scholars, language teaching scholars, NLP scholars

Website

https://info.clarin.dk/

Key achievements

Events
a. *CLARIN-DK's parliamentary data* [Poster presentation] (2023, November 23) - Sprogteknologisk Konference *2023*, co-organised by the Agency for Digital Government and Centre for Language Technology https://cst.ku.dk/kalender/sprogteknologisk-konference-2023/

Projects
a. Participation in ParlaMint II project

Resources (datasets, tools, services)
a. Van der Sluis, F., (2023). *Ensiwiki-2011 dataset for readability modelling*. CLARIN-DK-UCPH Centre Repository http://hdl.handle.net/20.500.12115/49

Publications
a. Navarretta, C., & Hansen, D. (2023). According to BERTopic, what do Danish parties debate on when they address energy and environment?. *ACM Cyber-Physical System*

Security Workshop. Proceedings of KONVENS 2023 Workshop on Computational Linguistics for Political and Social Sciences (CPSS 2023) (pp. 59-68). Association for Computational Linguistics. https://aclanthology.org/2023.cpss-1.6.pdf

b. Branco, A., Eskevich, M., Frontini, F., Hinrichs, E., de Jong, F., Kamocki, P., König, A., Lindén, K., Navaretta, C., Piasecki, M., Piperidis, S., Pitkänen, O., Simov, K., Skadina, I., Trippel, T., Witt, A., & Zinn, C. (2023). The CLARIN infrastructure as an interoperable language technology platform for SSH and beyond. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-023-09658-z

Other

a. New CLARIN agreement (2023-2024)

Estonia/CELR (Centre of Estonian Language Resources)	Center of Estonian Language Resources

Coordinator name and institution

Joshua Wilbur, University of Tartu

List of consortium partners (with centre types hosted)

1. University of Tartu
2. Tallinn University of Technology
3. Institute of the Estonian Language
4. Estonian Literary Museum

Together, these partners form one virtual C-centre, with plans to re-attain B-status in 2024.

Core user communities

Digital humanities scholars, linguists and language scholars; computational linguists, language technology researchers, NLP scholars; language technology developers

Website

https://keeleressursid.ee/en/

Key achievements

Events
   a. *Using CLARIN resources for corpus linguistics* - workshop by CLARIN ambassador Satu Saalasti at the MEDAL summer school in corpus linguistics at University of Tartu

Projects
   a. ParlaMint II (Estonia is included as of this version)

Resources (datasets, tools, services)
   a. In addition to Estonian, English, Finnish, Hungarian, Latvian, Norwegian, Russian, many Finno-Ugric languages (Komi, Komi Permyak, Udmurt, Hill Mari, Meadow Mari, Erzya, Moksha, Proper Karelian, Livvi Karelian, Ludian, Võro, Veps, Livonian, Northern Sami, Southern Sami, Inari Sami, Lule Sami, Skolt Sami, Mansi, Khanty) were added to machine translation tool (https://translate.ut.ee/); the corresponding language models and two

related datasets were published on hugging face
(https://huggingface.co/tartuNLP/smugri3_14-finno-ugric-nmt,
https://huggingface.co/datasets/tartuNLP/smugri-data,
https://huggingface.co/datasets/tartuNLP/smugri-flores-testset); see also this press
release (but neither CLARIN nor CELR is mentioned by name there)

b. Estonian National Corpus 2023 published

Publications

a. Kamocki, P., Linden, K., Puksas, A., & Kelli, A. (2023). EU Data governance act: Outlining a
potential role for CLARIN." In T. Erjavec, & M. Eskevich (Eds.), *Selected papers from the
CLARIN Annual Conference 2022,* 57–65. Linköping Electronic Conference Proceedings.
https://doi.org/10.3384/ecp198006

Finland/FIN-CLARIN	FIN-CLARIN

Coordinator name and institution

Krister Lindén (University of Helsinki)

List of consortium partners (with centre types hosted)

1. University of Helsinki
2. CSC – IT Centre for Scientific Computing, Ltd. (Kielipankki, type B-centre)
3. Kotus – Institute for the Languages of Finland
4. Aalto University
5. University of Eastern Finland
6. University of Jyväskylä
7. University of Oulu
8. University of Tampere
9. University of Turku
10. University of Vaasa

Core user communities

Digital humanities scholars, data scientists, cognitive scientists, Political science scholars, Social science scholars, Language and technology researchers, Speech technology researchers, Language technology developers, Software developers, Forensic linguists, corpus linguists, media scholars, Discourse scholars, Literary scholars, Translation scholars, Lexicographers, Translators, Conceptual and modern historians, Historians

Website

Language Bank of Finland: https://www.kielipankki.fi/news/
Consortium web page: http://www.helsinki.fi/fin-clarin

Key achievements

Events
   a. DHH23: Digital Humanities Hackathon - supported by CLARIN and DARIAH ERIC

Projects
   a. Upgrading of the **Korp server performance**.

Resources (datasets, tools, services)

    a. *FinnSentiment 1.1* - source material (27 000 sentences with sentiment annotation)
    b. *Finnish Dark Web Marketplace Corpus* **-** (findarc, 2 million posts)
    c. *Helsinki Corpus of English Texts, Early Modern English section* - source
    d. *Multimodal parallel corpus* - concordance demo in four Uralic languages

Collaborations (notable collaborations with other academic institutions, and/or GLAM sector or industry)

    a. *LAREINA* - project for developing speech recognition and synthesis modules for Finnish, Swedish and the Sámi languages in cooperation with industry

Publications

    a. Luukkonen, R. et al. (2023). FinGPT: Large generative models for a small language. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2710-2726. https://doi.org/10.48550/arXiv.2311.05640
    b. Rouhe, A., Grósz, T., & Kurimo, M. (2023). Principled comparisons for end-to-end speech recognition: Attention vs hybrid at the 1000-hour scale. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 32*, 623-638. https://doi.org/10.1109/TASLP.2023.3336517
    c. Jauhiainen, T., Zampieri, M., Baldwin, T., & Lindén, K. (2024). *Automatic language identification in texts*. Springer Nature.

Training materials

    a. *Corpus Linguistics and Statistical Methods* (5 ECTS) - Organized in the autumn and in the spring term, there are alternative study groups in English and in Finnish.
    b. *Introduction to Speech Analysis* (5 ECTS) - The course can be taken in English or in Finnish.
    c. *Data Clinic* (5 ECTS) - Online course, materials available in English. The general meetings will be organized online. The participants also work in small study groups with their peers. The groups may decide to meet either online or on the site of their preference.

Germany/CLARIN-D

Coordinator name and institution

Andreas Witt, Leibniz Institute for the German Language, Mannheim
Thorsten Trippel, University of Tübingen, deputy

List of consortium partners (with centre types hosted)

1. BBAW Berlin, CLARIN B-centre
2. EKUT Tübingen, Seminar für Sprachwissenschaft, CLARIN B-centre
3. GEI Braunschweig, Leibniz Institute for Educational Media | Georg Eckert Institute, CLARIN-C-centre
4. IDS Mannheim, CLARIN B/C-centre
5. LMU München, BAS, CLARIN B-centre
6. SAW Leipzig, CLARIN B-centre
7. UDS Saarbrücken, CLARIN B-centre
8. UFR, Centre for the Digital Foundation of Research in the Humanities, Social, and Educational Sciences (CEDIFOR), CLARIN-C-centre
9. UHH Hamburg, HZSK, CLARIN C-centre
10. UDUE, Duisburg, PolMine Projekct, CLARIN C-centre
11. UKÖ, Cologne, Language Archive Cologne, CLARIN C-centre
12. UST, Stuttgart, Institut für Maschinelle Sprachverarbeitung, CLARIN-C-centre

Core user communities

Linguists and language scholars, Language technology researchers, Speech technology researchers, Corpus linguists, Digital humanities scholars, Political science scholars, Social science scholars, Translation scholars, Information and communication scientists, Library researchers

Website

https://www.clarin-d.net/

Key achievements

In Germany many achievements are in conjunction with the National Research Data Infrastructure consortium Text+.

Events

a. *CLARIN Café: 'Do Chatbots Dream of Copyright?'* (2023, April 11) - Copyright in AI-generated Language Data https://www.clarin.eu/event/2023/clarin-cafe-do-chatbots-dream-copyright-copyright-ai-generated-language-data

b. *Outreach workshop on users depositing data* (2023, June 22) - Mannheim, Germany https://events.gwdg.de/event/372/

c. *Text+ Plenary 2023: Connecting People and Data* (2023, September 28-29) - Göttingen, Germany, https://events.gwdg.de/event/375/

d. *EuReCo Workshop* (2023, October 18) - Leuven, Belgium https://www.dropbox.com/scl/fi/dofkgy63z81depv7us4q3/EuReCo-Workshop-Description-Agenda_final2.pdf?rlkey=lrmx6yyzyrdcir98tcdtxsg9v&dl=0

e. *CLARIN and Libraries 2023* (2023 December 5) - Large Language Models and Libraries, Oslo, Norway https://www.clarin.eu/event/2023/clarin-and-libraries-2023-large-language-models-and-libraries

f. *Text+ Spring Meeting* (2024, March 12-13) - Frankfurt, Germany https://events.gwdg.de/event/607/

Resources (datasets, tools, services)

a. *LEXFCS* - Extending FCS for lexical resources

Publications

a. Branco, A., Eskevich, M., Frontini, F., Hajič, J., Hinrichs, E., Jong, F. de, Kamocki, P., König, A., Lindén, K., Navarretta, C., Piasecki, M., Piperidis, S., Pitkänen, O., Simov, K., Skadiņa, I., Trippel, T., Witt, A. & Zinn, C. (2023). The CLARIN infrastructure as an interoperable language technology platform for SSH and beyond. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-023-09658-z

b. Eckart, T., Herold, A., Körner E., & Wiegand F. (2023). A federated search and retrieval platform for lexical resources in Text+ and CLARIN. *Electronic Lexicography in the 21st Century: Proceedings of the eLex 2023 Conference.* Lexical Computing CZ s.r.o. https://elex.link/elex2023/wp-content/uploads/69.pdf

c. Hinrichs, E., & Trippel, T. (forthcoming). Text+: Concept and benefits for empirical researchers. *CLaDA-BG 2023 Conference: Language Technologies and Digital Humanities: Resources and Applications (LT⬚DH-RA)*. CEUR Workshop Proceedings.

d. Kahmann, C., Eckart, T., Körner, E., Helfer, F., & Schröder, C. (2023, December 12). *AI based methods for using text as data in the social sciences* [Conference presentation]. BERD Academy, Leipzig University, Leipzig, Germany. https://www.berd-nfdi.de/berd-academy/aitext-2023/

e. Kamocki, P., Lobin, H., Witt, A., & Wöllstein, A. (2024). Legal status and regulation of the German language in the Federal Republic of Germany'. *Current Issues in Language Planning*, 1–15. https://doi.org/10.1080/14664208.2023.2296787

Greece/clarin:el

Coordinator name and institution

Stelios Piperidis, ILSP/Athena RC
Maria Gavriilidou, ILSP/Athena RC, deputy

List of consortium partners (with centre types hosted)

1. ILSP/Athena RC (B-centre and NLP:EL K-centre)
2. NCSR DEMOKRITOS
3. GRNET
4. National Kapodistrian University of Athens
5. Aristotle University of Thessaloniki
6. University of the Aegean
7. Ionion University
8. Center for the Greek Language
9. Panteion University
10. National Center for Social Research
11. University of Crete
12. University of West Attica
13. University of Patras
14. Athens University of Economics and Business
15. University of Peloponnese

Core user communities

Language technology researchers, Digital humanities scholars, Social science scholars, Library researchers

Website

CLARIN:EL Portal: www.clarin.gr
CLARIN:EL Platform https://inventory.clarin.gr/
NLP:EL - CLARIN Knowledge Centre for Natural Language Processing in Greece
https://www.clarin.gr/el/kcentre

Key achievements

Events
a. Organisation of a series of teaching and training events:
   - *Webinar1: Exploring CLARIN* (2023, November 30) - Introduction and familiarization of participants with data search and retrieval through CLARIN and CLARIN:EL (Kanella Pouli, ILSP/ATHENA RC) https://www.clarin.gr/en/node/1398

28

- *Webinar 2: How did we get to ChatGPT and what we can do with it* (2024, March 7) - Stergios Chatzikyriakidis, Professor of Computational Linguistics, University of Crete https://www.clarin.gr/en/content/clarinel-webinar-series
- *Webinar 3: ParlaMint Parliamentary corpora: CLARIN ERIC Flagship Project* (2024, March 11) - Maria Gavriilidou, Researcher, ILSP/ATHENA RC https://www.clarin.gr/en/content/clarinel-webinar-series
- *2 hands-on labs* (2024, January 26 and March 3) - Introduction, familiarization and use of NLP tools and services of CLARIN:EL (Juli Bakagianni, ILSP/ATHENA RC) https://www.clarin.gr/el/content/praktika-ergastiria-clarinel-exoikeiosi-me-ta-ergaleia-epexergasias-fysikis-glossas-tis

b. *Organisation of National project DIGITAL LANDSCAPE: Final event* (2024, March 13)

- Presentation of results https://digital-landscape.gr/2024/03/01/%ce%ba%ce%b1%cf%84%ce%b1%ce%bb%ce%b7%ce%ba%cf%84%ce%ae%cf%81%ce%b9%ce%b1-%ce%b5%ce%ba%ce%b4%ce%ae%ce%bb%cf%89%cf%83%ce%b7-%cf%88%ce%b7%cf%86%ce%b9%ce%b1%ce%ba%ce%bf%cf%8d-%cf%84%ce%bf%cf%80%ce%af/

Projects

a. *Digital Landscape* [successfully completed] https://digital-landscape.gr/2024/03/01/%ce%ba%ce%b1%cf%84%ce%b1%ce%bb%ce%b7%ce%ba%cf%84%ce%ae%cf%81%ce%b9%ce%b1-%ce%b5%ce%ba%ce%b4%ce%ae%ce%bb%cf%89%cf%83%ce%b7-%cf%88%ce%b7%cf%86%ce%b9%ce%b1%ce%ba%ce%bf%cf%8d-%cf%84%ce%bf%cf%80%ce%af/
b. *ParlaMint II* https://www.clarin.eu/parlamint
c. *LDS: Common European Language Data Space* (2023, January - 2025, December) - https://www.athenarc.gr/en/institute-language-and-speech-processing/projects/lds-common-european-language-data-space
d. *DeMoLiSH Detecting, Monitoring and Linking online Hate Speech and offline Hate Crimes* (2023, October - 2025, August) - https://www.athenarc.gr/en/institute-language-and-speech-processing/projects/detecting-monitoring-and-linking-online-hate

Resources (datasets, tools, services)

a. *ParlaMint-GR* - the Greek part of the ParlaMint corpus
b. 220 new resources added (corpora and lexical resources)
c. *Meltemi* - the first Greek Large Language Model
d. New services integrated in *CLARIN:EL* - Verbal aggression analysis in Twitter, Greek and English
e. New user dashboard.

Publications

a. Gavriilidou, M., Gkoumas, D., Prokopidis, P., Papavassiliou, V., & Piperidis, S. 2023. *The ParlaMint-GR corpus: Annotated Greek parliamentary proceedings.* 16th International

Conference on Greek Linguistics, *Thessaloniki, Greece*.
https://www.new.enl.auth.gr/icgl16/

b. Gavriilidou, M., Gkoumas, D., Prokopidis, & Piperidis, S. 2023. Investigating Political Ideologies through the Greek ParlaMint corpus. *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*.

c. Branco, A., Eskevich, M., Frontini, F., Hajič, J., Hinrichs, E., de Jong, F., Kamocki, P., König, A., Lindén, K., Navarretta, C., Piasecki, M., Piperidis, S., Pitkänen, O., Simov, K., Skadiņa, I., Trippel, T., Witt, A., & Zinn, C. (2023). The CLARIN infrastructure as an interoperable language technology platform for SSH and beyond. *Language Resources and Evaluation Journal.* https://doi.org/10.1007/s10579-023-09658-z

d. Gavriilidou et al. (2023). The CLARIN:EL infrastructure. *CLARIN Annual Conference 2023*. https://office.clarin.eu/v/CE-2023-2328_CLARIN2023_ConferenceProceedings.pdf

Other

a. One new member joined the CLARIN:EL network (U. of Peloponnese)

Hungary/HunCLARIN	HUNCLARIN

Coordinator name and institution

Tamás Váradi, Hungarian Research Centre for Linguistics

List of consortium partners (with centre types hosted)

1. Hungarian Research Centre for Linguistics – coordinator
2. Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics
3. HUN-REN Institute for Computer Science and Control
4. HUN-REN Research Centre for Natural Sciences, Institute of Cognitive Neuroscience and Psychology
5. HUN-REN Centre for Social Sciences, Institute for Political Science, poltextLAB
6. University of Szeged
7. Pázmány Péter Catholic University

Core user communities

Linguists and language scholars, Software developers, NLP scholars, Psychologists, Sociologists

Website

www.clarin.hu

Key achievements

Events

a. *Researcher's Night, PULI: The Hungarian Artificial Intelligence* (2023, September 29) - Now you instruct! https://app.kutatokejszakaja.hu/esemenyek/nyelvtudomanyi-kutatokozpont/puli-a-magyar-mesterseges-intelligencia-most-te-parancsolsz

b. *ChatGPT: widely used, yet not fully understood* [Chat GPT: tömegek használják, de senki sem érti (meg)](2023, September 3) - Roundtable discussion featuring Gábor Prószéky, Tamás Váradi, and members of the Language Technology Research Group at the Research Institute for Linguistics of the Hungarian Academy of Sciences (NYTK) about the creation, operation, strengths, and limitations of the ChatGPT system, as well as clearing up some misconceptions.

Projects

a. Supporting the Digital Sustainability of the Hungarian Language: The current project focuses on research and development within four subprojects, in which the Research Institute for Linguistics, recognized as an Excellent Research Place by the Hungarian Academy of Sciences (MTA), plays a leading and unique role. These are the following:
   - *Hungarian National Corpus 3.0*
   - *Spelling Advisory Portal 2.0*
   - *Digitization of the Archival Card Collection of the Comprehensive Dictionary of the Hungarian Language*
   - *Analyzed Corpus and Knowledge Base of the Ob-Ugric Languages*

b. Digital Support of the Hungarian Language in the Service of Hungarian Scholarship: In general, it aids scientific communication through text mining methods and, within this scope, provides significant support for Hungarian scholarship by processing and making Hungarian-language scientific publications accessible. Within the project, we primarily plan to extract modern, born-digital, peer-reviewed content that can be effectively used as training material. The language technology tools developed within the project will thus enable the improvement and enrichment of the REAL repository's content and enhance the repository services provided to users (researchers).

Resources (datasets, tools, services)

a. *Hungarian ParlaMint II* - The newly built Hungarian corpus of ParlaMint II focuses on the minutes of the Hungarian National Assembly between May 2014 and June 2022 and can be considered a major improvement from the Hungarian corpus of ParlaMint I.
b. *PULI GPT 3SX* - The first Hungarian-only GPT-3 base model.
c. *PULI GPTrio* - A trilingual (Hungarian, English, Chinese) GPT-3 model.
d. *ParancsPULI* - Hungarian-language instruction follower fine-tuned PULI-GPTrio.

Collaborations (notable collaborations with other academic institutions, and/or GLAM sector or industry)

Library and Information Centre of the Hungarian Academy of Sciences

Publications

a. Yang, Z. G., Dodé, R., Héja, E., Laki, L. J., Ligeti-Nagy, N., Madarász, G., Váradi, T. (2024). ParancsPULI: Az utasításkövető PULI-modell. In G. Berend, G. Gosztolya, & V. Vincze (Eds.), *XX. Magyar Számítógépes Nyelvészeti Konferencia* (pp. 61-72). Szegedi Tudományegyetem.
b. Yang, Z. G., Szlávik, S., & Ligeti-Nagy, N. (2024). Magyar nyelvű utasításkövető korpusz építése Stanford Alpaca promptok fordításával és lokalizálásával. In G. Berend, G. Gosztolya, & V. Vincze (Eds.), *XX. Magyar Számítógépes Nyelvészeti Konferencia* (pp. 243-255). Szegedi Tudományegyetem.
c. Ligeti-Nagy, N., Héja, E., Takács, D., Yang, Z. G., & Váradi, T. (2023). Hát te mekkorát nőttél! – A HuLU első életéve új adatbázisokkal és webszolgáltatással. In G. Berend, G.

Gosztolya, & V. Vincze (Eds.), *XIX. Magyar Számítógépes Nyelvészeti Konferencia, MSZNY-2023* (pp. 217-230). Szegedi Tudományegyetem TTIK, Informatikai Intézet.

d. Ligeti-Nagy, N., Dodé, R., Jelencsik-Mátyus, K., Varga, Z., Héja, E., & Váradi, T. (2023). A new ParlaMint corpus for Hungarian 30m tokens of annotated parliamentary data. In G. Berend, G. Gosztolya, & V. Vincze (Eds.), *XIX. Magyar Számítógépes Nyelvészeti Konferencia, MSZNY-2023* (pp. 433–446). Szegedi Tudományegyetem TTIK, Informatikai Intézet.

e. Yang, Z. G., Dodé R., Ferenczi G., Héja E., Kőrös, A., Laki, L. J., Ligeti-Nagy, N., Jelencsik-Mátyus, K., Vadász, N., Váradi, T. (2023). Jönnek a nagyok! BERTLarge, GPT-2 és GPT-3 nyelvmodellek magyar nyelvre. In G. Berend, G. Gosztolya, & V. Vincze (Eds.), XIX. Magyar Számítógépes Nyelvészeti Konferencia, MSZNY-2023 (pp. 247–262). Szegedi Tudományegyetem TTIK, Informatikai Intézet.

Iceland/CLARIN-IS

Coordinator name and institution

Starkaður Barkarson (The Árni Magnússon Institute for Icelandic Studies)

List of consortium partners (with centre types hosted)

1. The Árni Magnússon Institute for Icelandic Studies (B-centre)
2. The University of Iceland
3. Reykjavik University
4. National and University Library of Iceland
5. National Archives of Iceland
6. The Icelandic Language Council
7. The Icelandic National Broadcasting Service
8. Almannarómur - Centre for Language Technology

Core user communities

Linguists and language scholars, Language technology researchers

Website

www.clarin.is

Key achievements

Projects

a. We participated in a project launched by the Faculty of Education in creating a *Corpus of Icelandic Academic Vocabulary* and compiling *The Icelandic Academic Word List.*

Resources (datasets, tools, services)

a. New and updated version of KORP published.

Other

a. CLARIN Center at the Arni Magnusson Institute for Icelandic Studies became a B-centre.
b. K-centre for Icelandic established

Italy/CLARIN-IT

Coordinator name and institution

Monica Monachini, Institute for Computational Linguistics 'A. Zampolli' – National Research Council (ILC-CNR)

List of consortium partners (with centre types hosted)

1. Institute for Computational Linguistics "A Zampolli" – CNR, Pisa (ILC4CLARIN, type B and co-coordinator of DipText KC, type K.
2. Center for Comparative Studies "I Deug-Su" – Dept. of Philology and Literary Criticism – University of Siena (DFCLAM-UNISI)
3. Institute for Applied Linguistics – Eurac Research, Bolzano (ERCC, type C)
4. Bruno Kessler Foundation (FBK), Trento
5. Archival and Bibliographical Superintendence of Tuscany (SAB-TOS), Firenze
6. Dept of Electrical Engineering and Information Technology - (DIETI-UNINA)
7. Interdepartmental Research Center "URBAN/ECO" - University of Naples Federico II (URBAN/ECO-UNINA), Napoli
8. Catholic University of the Sacred Heart (UNICATT), Interdisciplinary Research Center "CIRCSE", Milano
9. Parma University
10. Padua University
11. University Language Center, University of Ferrara
12. The University of Naples L'Orientale
13. The RUT Foundation

Core user communities

Digital humanities scholars, Linguists and language scholars, Language teaching scholars, Classical scholars, NLP scholars, Archivists

Website

http://www.clarin-it.it

https://diptext-kc.clarin-it.it/

Key achievements

Events

a. *CNR-ILC-DipText KC* (2022, November 22) - Linguistic annotation with IT tools: An Update and training activity for teachers, part of the Updating and Training Activities of the 'Latin Language of Science' initiative. https://diptext-kc.clarin-it.it/2023/01/19/concluded-the-updating-and-training-activities-of-the-latin-language-of-science-initiative/
b. *La memoria digitale* (2023, June 5-7) - Forme del testo e organizzazione della conoscenza. XII Conference of AIUCD (co-organised with AIUCD Association & UniSiena) https://www.unisi.it/unisilife/eventi/la-memoria-digitale-forme-del-testo-e-organizzazione-della-conoscenza
c. *Digital tools for humanists* [Summer School, Pisa] (2023, June 13-22) -Supported by CLARIN. Focus of the 2023 edition: tools and applications dealing mostly with Digital Public History
d. *CLARIN Café: A new CLARIN Resource Family for Lexical Semantic Change Research* (2023, July 5) - Co-organised CLARIN-IT-CLARIN-UK https://www.clarin.eu/event/2023/clarin-cafe-new-clarin-resource-family-lexical-semantic-change-research
e. *Infrastrutture e Risorse per la Ricerca Linguistica in Italia* (2023, November 11) - organized in the context of the national PNRR IR H2IOSC [Humanities and cultural Heritage Italian Open Science Cloud] project to discuss the role of RIs and elicit community needs and desiderata; https://www.clarin-it.it/it/content/evento-h2iosc-pisa

Projects

a. *MUSAE: Music and Arts Empowerment* (Call: AFAM – NextGenerationEU)
The focus is to create a circuit of exchanges, educational and research activities, events and artistic productions along the two main strands of traditional music (a common element in all countries and a common ground for encounters and artistic creation), and of some Italian excellences that are still highly attractive and internationally recognizable (opera, accordion), enhanced through the modern tools of digitization.
b. *OSTrails: Open Science Plan-Track-Assess Pathways* (Call: HORIZON-INFRA-2023-EOSC-01 - Project number: 101130187).
OSTrails aims to advance processes and instruments for Planning, Tracking, and Assessing scientific knowledge production beyond state-of-the art, working with various national and thematic contexts, improving existing infrastructure, and connecting key components.
c. *ECHOES: European Cloud for Heritage OpEn Science* (Call: HORIZON-CL2-2023-HERITAGE-ECCCH-01 – Project number: 101157364).
ECHOES aims to create the European Collaborative Cloud for Cultural Heritage (ECCCH) as a shared platform for heritage professionals and researchers to access data, innovative scientific and training resources and advanced digital tools co-developed by the heritage community according to their specific needs.
d. *Testo: Translating, Encoding, Sharing The Origins*. From the Littera Florentina to an open access Italian translation of Justinian's Digest. (Call PRIN PNRR 2022 – Project number:

P20224NJLK). *The project intends to create a complete and updated Italian translation of the text of the Digest.*

e. *Roads: Roads to Oral Archives Development and Sustainability* (Call PRIN 2022 – Project number: 2022TSXM2P). *ROADS deals with oral archives produced by academic researchers in different fields (e.g. oral history and linguistics). It envisages a framework (under the CLARIN-IT umbrella) capable of ensuring long-term preservation, security, standardization, compliance with ethical principles and legal issues, and sustainability, according to the FAIR principles.*

Resources (datasets, tools, services)

a. Augustine of Hippo. n.d. *Confessions* [Corpus]. The Latin Library. http://www.thelatinlibrary.com/august.html [The original text was lemmatized and PoS tagged with the UDPipe tool (using the PROIEL trained model).]

b. TrAVaSI_VoDIM [Corpus]. (2018). In C. Marazzini & L. Maconi (Eds.), *Vocabolario Dinamico Dell'Italiano Moderno*. [The TrAVaSI_VoDIM Corpus is a sample of the corpus built for the Vocabolario Dinamico Dell'Italiano Moderno (VoDIM), gathering Italian texts from 1861 to the present day, after the Unification of Italy. TrAVaSI_VoDIM is balanced and representative of different prose domains (art, gastronomy, law, newspapers, literature, popular fiction, science), for a total of about 21.000 tokens. TrAVaSI_VoDIM is morpho-syntactically annotated and lemmatized.]

c. Zanola, M. T., Dankova, K., Grimaldi, C., & Serpente, A. (2023). *Pan-Latin Lexicon of Collars and Sleeves in Fashion and Costume*. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, Pisa. Pan-Latin Lexicon of Collars and Sleeves in Fashion and Costume (cnr.it) [Lexical Conceptual Resource. The Pan-Latin Lexicon of Collars and Sleeves in Fashion and Costume, developed within the Pan-Latin Terminology Network (REALITER), aims at collecting the main terms designating collars and sleeves in fashion and costume. The Lexicon intends to give a multilingual (Italian, Catalan, Spanish, French, Portuguese, English) terminological description in this field.]

d. Imparato, S. C., et al. (2023). *ItaASD: Italian speech corpus Autism Spectrum Disorder*. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, Pisa. http://hdl.handle.net/20.500.11752/OPEN-990 [This is a corpus of semi-spontaneous speech produced by 34 children between 6 and 13 years of age, residents in the Campania region of Italy. Half of the participating children were diagnosed with high-functioning Autism Spectrum Disorder, and the other half were neurotypical children matched for age, gender, and geographical origin. Processed data (i.e., tables of acoustic/rhythmic/lexical/syntactic values, with the name of the speakers masked through an alphanumeric acronym to ensure anonymity) are available from the contact person upon reasonable request.]

e. *EpiLeXo* [ToolService]. EpiLexO is a user-friendly web application for the creation and editing of an integrated system of language resources for ancient fragmentary languages centered on the lexicon, in compliance with current digital humanities and Linked Open Data principles. EpiLexo allows for the editing of lexica with all relevant cross-references: for their linking to their testimonies, as well as to bibliographic information and other (external) resources and common vocabularies.

Collaborations (notable collaborations with other academic institutions, and/or GLAM sector or industry)
   a. Conservatory of Cosenza - PNRR AFAM Project
   b. RUT Foundation - Intercultural dialogue and social innovation - Social dimension of scientific research
   c. Ferrara University - development of training modules
   d. Tuscany Superintedence - archives' preservation.

Publications
   a. Degl'Innocenti, E., Monachini, M., Bucciero, A., Pasini, E., Fanini, B., & Frontini, F. (Forthcoming). H2IOSC: Humanities and Heritage Open Science Cloud - Presentation of the PNRR H2IOSC project, AIUCD 2023 - Book of Extended Abstracts. *Quaderni di Umanistica Digitale*.
   b. Monachini, M., & Frontini, F. (2023). Infrastrutture digitali per le scienze umane e sociali. In F. Ciotti (Ed.), *Digital Humanities: Metodi, Strumenti, Saperi* (pp. 197–213). Carocci Editore.
   c. Calamai S., Piccardi D., Pretto N., Candeo G., Stamuli M. F. & Monachini M. (2022). Not just paper: Enhancement of archive cultural heritage. In D. Fišer & A. Witt (Eds.), *CLARIN The Infrastructure for Language Resources, Vol. 1*. De Gruyter.
   d. Venuti M. C., Del Grosso A. M., Boschetti F., Tessarolo L., Prontera A., Bovet D., Cattaneo G. & Melis V. (2023). La 'Galassia MQDQ': Un concetto di filologia tradizionale, digitale, sostenibile. *Magazèn 4*, 71-120. http://doi.org/10.30687/mag/2724-3923/2023/07/003
   e. Frontini, F., Romary, L., & Khan, A. F. (2023). ISO LMF 24613-6: A Revised Syntax Semantics Module for the Lexical Markup Framework. *Proceedings of LDK 2023 – 4th Conference on Language, Data and Knowledge, Vienna, Austria*.

Training materials
   a. Del Grosso, A. M. (2023, October 19-20). *Processing of XML Documents.* Training School 2023 on Digital Text Processing. https://diptext-kc.clarin-it.it/digital-text-processing-2023/
   b. Monachini, M. & Frontini, F. (2023, June 16). *Research Infrastructures supporting FAIR and Open data, tools, practices in the humanities and social sciences: The case of CLARIN*. Lesson at the SUMMER SCHOOL "DIGITAL TOOLS FOR HUMANISTS", Pisa. https://digitaltools.labcd.unipi.it
   c. Sprugnoli, R. (2023, June 22). *Natural Language Processing methods*. Seminar at the SUMMER SCHOOL "DIGITAL TOOLS FOR HUMANISTS", Pisa. https://digitaltools.labcd.unipi.it, https://digitaltools.labcd.unipi.it/risorse/#Sprugnoli, https://drive.google.com/drive/folders/1lCigkmAesQXWp8REg5aZxQDjzLDV_RT2
   d. Del Fante, D. (2023, June 16). *Clarin-IT Using the CLARIN infrastructure for DH research, the example of ParlaMint* [Presentation]. Summer School Digital Humanities and Digital Communication: AI and (new) literacies, Modena, June 12th-16th. https://clarin-it.it/en/content/clarin-it-unimore-summer-school-digital-humanities-2023
   e. Boschetti, F. (2023, January 19). *Semi-automatic transcription with eScriptorium (CNR-ILC-CoPhiLab)* [Training activity fo students]. Updating and Training Activities of the 'Latin Language of Science' initiative. https://diptext-kc.clarin-it.it/2023/01/19/concluded-the-updating-and-training-activities-of-the-latin-language-of-science-initiative/

Latvia/CLARIN-LV	CLARIN-LV

Coordinator name and institution

Inguna Skadiņa, Institute of Mathematics and Computer Science, University of Latvia (IMCS UL)
Ilze Auziņa, IMCS UL, deputy

List of consortium partners (with centre types hosted)

1. Institute of Mathematics and Computer Science of the University of Latvia (B-Centre)
2. University of Latvia
3. Institute of Literature, Folklore and Art of the University of Latvia
4. Rīga Stradiņš University
5. Liepaja University
6. Rezekne Academy of Technologies
7. National Library of Latvia

Core user communities

Digital humanities scholars, Media scholars, Information and communication scientists, Linguists and language scholars, Translation scholars, Software developers

Website

http://clarin.lv/
CLARIN-LV repository: https://repository.clarin.lv/

Key achievements

Events

a. *Recent Language Resources and Tools for Digital Humanities* (Jaunākie valodas resursi un rīki digitālajām humanitārajām zinātnēm) [Conference] (2023, September 12) - https://clarin.lv/lv/clarin-latvija-seminari/99-konferencei-jaunakie-valodas-resursi-un-riki-digitalajam-humanitarajam-zinatnem
b. *Creation and Usage of Speech Corpora* (Runas korpusu izveide un to izmantojums) [Seminar] (2023, May 18) - https://clarin.lv/lv/clarin-latvija-seminari/94-seminars-runas-korpusu-izveide-un-to-izmantojums

Projects

a. *Language Technology Initiative* - https://www.vti.lu.lv/en/
b. *LATE* - Research on Modern Latvian Language and Development of Language Technology https://www.digitalhumanities.lv/projects/vpp-late/

c. *DHELI* - Towards Development of Open and FAIR Digital Humanities Ecosystem in Latvia https://www.digitalhumanities.lv/projects/DHELI-en/
d. *COST Action UniDive* – EC Project No. CA21167 (cooperation with 250 researchers in 36 COST countries + 5 more countries)
e. Participation in *ParlaMint 2* project.

Resources (datasets, tools, services)

a. *National Corpus Collection* - https://korpuss.lv/en/
b. Spektors, A., et al. (2023). *Tēzaurs.lv 2023 (*Autumn Edition). CLARIN-LV digital library at IMCS. University of Latvia. http://hdl.handle.net/20.500.12574/92
c. Rituma, L., Pretkalniņa, L., Saulīte, B., Nešpore-Bērzkalne, G., & Grūzītis, N. (2023). *LVTB - Latvian Treebank v2.13* (2023-11-15). CLARIN-LV digital library at IMCS. University of Latvia. http://hdl.handle.net/20.500.12574/91
d. Jērāne, S., et al. (2023). *Dictionary of Contemporary Latvian Language (MLVV)* (2023-07-07). CLARIN-LV digital library at IMCS. University of Latvia. http://hdl.handle.net/20.500.12574/88
e. Levāne-Petrova, K., Darģis, R., Pokratniece, K., & Lasmanis, V. J. (2023). *Balanced Corpus of Modern Latvian (LVK2022)*. CLARIN-LV digital library at IMCS. University of Latvia. http://hdl.handle.net/20.500.12574/84

Collaborations (notable collaborations with other academic institutions, and/or GLAM sector or industry)

a. Through DHELI and LATE projects
b. *Balsu Talka* - crowdsourcing speech data https://balsutalka.lv/

Publications

a. Branco, A., Eskevich, M., Frontini, F., Hajic, J., Hinrichs, E., de Jong, F., Kamocki, P., König, A., Lindén, K., Navarretta, C., et al. (2023). The CLARIN infrastructure as an interoperable language technology platform for SSH and beyond. *Language Resources and Evaluation.*
b. Skadina, I., Auzina, I., Valkovska, B., & Gruzitis, N. (2023). Language Report Latvian. *European Language Equality*.
c. Grasmanis, M., Paikens, P., Pretkalnina, L., Rituma, L., Strankale, L., Znotins, A., Gruzitis, N. (2023). Tēzaurs.lv – the experience of building a multifunctional lexical resource. *Proceedings of the eLex 2023 conference, Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography*. Lexical Computing CZ s.r.o.
d. Rituma, L., Nespore-Berzkalne, G., Saulite, B., Pretkalnina, L. (2023). Vārdkopas analogi „Latviešu valodas sintaktiski marķētajā korpusā". *Valoda: nozīme un forma (Analogue of subordinate phrase in Latvian Treebank)*, 156-173.

Training materials

a. Are being created through Language Technology Initiative, information on courses: Study modules (lu.lv)

Other

a. B-centre status and CTS obtained.

Lithuania/CLARIN-LT

Coordinator name and institution

Jurgita Vaičenonienė, Vytautas Magnus University

List of consortium partners (with centre types hosted)

1. Vytautas Magnus University (coordinating institution, C-centre)
2. Kaunas University of Technology
3. Vilnius University
4. Mykolas Romeris University
5. Baltic Institute of Advanced Technology

Core user communities

Digital humanities scholars, Linguists and language scholars, Translation scholars, Translators, Social science scholars, Software developers

Website

http://clarin-lt.lt/
https://sitti.vdu.lt/en/

Key achievements

Events

a. *Initiatives supporting Ukrainian language resources and language resource and technology development* [presentation] (2023, May 22) - Seminar Scientific Data about Society in Ukraine, Ukraine in European Research Infrastructures in the Social Sciences and Humanities https://www.esfri.eu/webinar-data-society-in-ukraine
b. *CLARIN Resources and Language Analysis Tools: Application in Educational Settings* [Pre-conference workshop] (2023, June 8) - International scientific conference 'Sustainable Multilingualism 2023' https://sites.google.com/view/sm2023/programme
c. *Seminar for Ukrainian Knowledge centre CLARIN: Open Access Resources, Tools and Services for Language Research* (2023, December 15) - https://uacorpus.org/k-centre/index.html#section_2

Resources (datasets, tools, services)

a. New resources uploaded to CLARIN-LT repository: https://clarin.vdu.lt/xmlui/?locale-attribute=lt

Collaborations (notable collaborations with other academic institutions, and/or GLAM sector or industry)

a. Reviewing research proposals for Supporting as Research Reviewer the EURIZON Fellowship Call *Remote Research Grants for Ukrainian researchers* https://www.eurizon-project.eu/news/calls/call_for_applications/

Publications

a. Kamocki, P., Linden, K., Puksas, A., Kelli, A. (2022). EU data governance act: Outlining a potential role for CLARIN. *Selected papers from the CLARIN Annual Conference*. https://ecp.ep.liu.se/index.php/clarin/issue/view/79
b. CLARIN news report: https://www.clarin.eu/blog/initiatives-supporting-ukrainian-researchers-and-language-resource-and-technology-development

Training materials

a. CLARIN-LT content continuously integrated into study curricula.

Other

a. CLARIN-LT plan for 2023-2027 sent for approval to the Research Council of Lithuania.

The Netherlands/CLARIAH-NL

Coordinator name and institution

Gijsbert Rutten (Leiden University)
Menzo Windhouwer (HuC-KNAW), deputy

List of consortium partners (with centre types hosted)

1. Huygens ING, Amsterdam (C-centre)
2. Utrecht University, Utrecht
3. Dutch Language Institute, Leiden (B-centre)
4. Netherlands Institute for Sound and Vision, Hilversum (T-centre)
5. Meertens Institute, Amsterdam (B-centre)
6. Data Archiving and Networked Services, The Hague (C-centre)
7. Radboud University, Nijmegen (C-centre)
8. Free University, Amsterdam
9. University of Groningen, Groningen
10. Fryske Akademy, Ljouwert
11. Leiden University, Leiden
12. University of Amsterdam, Amsterdam
13. International Institute of Social History, Amsterdam
14. National Library, The Hague

Core user communities

Linguists and language scholars, Literary scholars, Historians, Political science scholars, Theologists, Social and economic historians.

Website

http://www.clariah.nl; CLARIAH-PLUS is a national infrastructure project that contributes to CLARIN and DARIAH.

Key achievements

Events
a. *Annual Conference* (2023, November 30) - https://www.clariah.nl/events/clariah-annual-conference-2023

Projects

    a. Follow-up project *SSHOC-NL (Social Science and Humanities Open Cloud for the Netherlands)* [awarded] - https://odissei-data.nl/en/2023/02/nwo-roadmap-grant-for-digital-infrastructure-social-sciences-and-humanities/

Resources (datasets, tools, services)

    a. beta version of the CLARIAH-NL vocabulary registry
    b. first version of CLARIAH-NL central catalog: INEO
        - fed by a tools and data harvest filtered on FAIRness

Collaborations (notable collaborations with other academic institutions, and/or GLAM sector or industry)

    a. The Koninklijke Bibliotheek (KB, Royal Library) has been part of the national infrastructure for years. The Netwerk Digitaal Erfgoed (NDE, Network Digital Heritage) participates in SSHOC-NL.

Publications

    a. Lee, A., Bessell, N., van den Heuvel, H. et al. (2023). The DELAD initiative for sharing language resources on speech disorders. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-023-09701-z

Norway/CLARINO

Coordinator name and institution

Koenraad De Smedt, University of Bergen

List of consortium partners (with centre types hosted)

1. UiB University of Bergen (CLARINO Bergen Centre, type B and K-centre for Treebanking)
2. UiO University of Oslo (Text Laboratory, type C)
3. UiT The Arctic University of Norway (TROLLing, type C)
4. SIKT Norwegian Agency for Shared Services in Education and Research (K-centre for Data Management)
5. NHH Norwegian School of Economics, Bergen
6. NB National Library of Norway (Språkbanken, type C)

Core user communities

Linguists and language scholars, Philosophers, Linguists and language scholars, Digital humanities scholars, Historians, Scholars of law, Social science scholars, Political science scholars, Lexicographers, Software developers, Language technology developers

Website

http://clarin.w.uib.no

Key achievements

Resources (datasets, tools, services)
a. *Norwegian Media Corpus* - https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-4/
b. *Abkhazian National Corpus* - https://clarino.uib.no/abnc/page
c. *Norwegian Dialect Change Corpus* - https://tekstlab.uio.no/nota/scandiasyn/
d. *Homer Corpus*
e. *Nature Corpus*
f. Nynorsk Corpus - https://tekstlab.uio.no/norsk/bokmaal/english.html
g. *GNC Rustaveli Parallel Corpus* [with Abkhaz translation]
h. *Universal Dependencies 2.12*

Collaborations (notable collaborations with other academic institutions, and/or GLAM sector or industry)

a. *Nonmanual* (sign language)
b. *MeNoTa* (Medieval Nordic Text Archive)
c. *ALAN* (learner language)
d. *Talebanken* (Norwegian dialects)
e. *KoPeSC* (Pomor speech)

Publications

a. Breder Birkenes, M., Gunnarsønn Bagøien Johnsen, L., Kåsen, A. (2023). NB DH-LAB: A corpus infrastructure for social sciences and humanities computing. *Proceedings of the CLARIN Annual Conference 2023*.
b. Rosén, V. (n.d.). LFG Treebanks. *The Handbook of Lexical-Functional Grammar*. Language Science Press.

Training materials

a. Johansen, H., Kyrkjebø, R., Meurer, P. (n.d.). *Training videos (in Norwegian) on searching the VLO, depositing language resources, use of Corpuscle and of Glossa*.

Poland/CLARIN-PL	**CLARIN** Common Language Resources and Technology Infrastructure
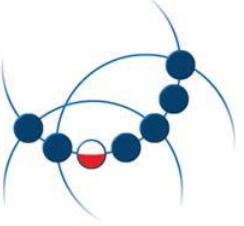
Coordinator name and institution

Maciej Piasecki, Wrocław University of Science and Technology

List of consortium partners (with centre types hosted)

1. Wrocław University of Science and Technology (hosted Centres: B and K)
2. University of Wrocław
3. Institute of Computer Science Polish Academy of Sciences
4. Institute of Slavic Studies Polish Academy of Sciences
5. University of Lodz
6. Polish-Japanese Academy of Information Technology

Core user communities

Digital humanities scholars, Literary scholars, Linguists and language scholars, Sociologists, Economists, Historians, Media scholars, Psychologists, NLP scholars, Software developers.

Website

https://clarin-pl.eu/index.php/en/home/
https://clarin.biz/

Key achievements

Events
a. *CLARIN-PL in Research Practice* [Workshop] (2023, May 11-12) - 60 participants, the Institute of Computer Science of the Polish Academy of Sciences, Warsaw.
b. *Competition* for the best publication involving the use of CLARIN-PL infrastructure (2023, May 1 - August 31) - 10 submissions.

Publications

a. Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, L., Wojtasik, K., Woźniak, & S., Kazienko,

P. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion*. https://doi.org/10.1016/j.inffus.2023.101861

b. Rudnicka, E. Grabowski, L., Piasecki, M., & Naskręt, T. (2023). In search of gaps between languages and wordnets: The case of Polish-English WordNet. *International Journal of Lexicography, 36*(1), 68–92. https://doi.org/10.1093/ijl/ecac005

c. Czeranowska, O., Chlasta, K., Miłkowski, P., Grabowska, I., Kocoń, J., Hwaszcz, K., Wieczorek, J., & Jastrzębowska, A. (2023). Migrants vs. stayers in the pandemic: A sentiment analysis of Twitter content. *Telematics and Informatics Reports 10*. https://doi.org/10.1016/j.teler.2023.100059

d. Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, N., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Calzada Pérez, M., de Macedo, L. D., Navarretta, C., Luxardo, G., Coole, M., Rayson, P., Morkevičius, V., Krilavičius, T., Darģis, R., Ring, O., van Heusden, R., Marx, M., & Fišer, D. (2023). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation 58*, 415–448. https://doi.org/10.1007/s10579-021-09574-0

e. Pęzik, P., Mikołajczyk, A., Wawrzyński, A., Żarnecki, F., Nitoń, B., & Ogrodniczuk, M. (2023). Transferable keyword extraction and generation with text-to-text language models. In J. Mikyška, C. de Mulatier, M. Paszynski, V. V. Krzhizhanovskaya, J. J. Dongarra, P. M.A. Sloot (Eds.), *Computational Science – ICCS 2023, 23rd International Conference, Prague, Czech Republic, Proceedings, Part II*. Lecture Notes in Computer Science 14074, 398–405. https://doi.org/10.1007/978-3-031-36021-3_42

Other

a. Securing funding for maintenance (2023-2024), promotion and infrastructure development from the Polish Ministry of Education and Science fund
b. approx. 450 identified scientific users of the infrastructure; 139 identified citations and references to CLARIN-PL tools and resources; 53 publications authored by CLARIN-PL team members.

Portugal/PORTULAN CLARIN

Coordinator name and institution

António Branco, University of Lisbon

List of consortium partners (with centre types hosted)

Proponent partners:
1. Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal (coordinator)
2. Faculdade de Letras da Universidade de Lisboa, Lisbon, Portugal
3. Universidade de Évora, Évora, Portugal

Implementation partners:
1. Centro de Estudos de Linguística Geral e Aplicada (CELGA-ILTEC), Coimbra, Portugal
2. Centro de Estudos Humanísticos (CEHUM), Braga, Portugal
3. Centro de Estudos Filosóficos e Humanísticos (CEFH) da Universidade Católica, Braga, Portugal
4. Centro de Investigação do Núcl. Est. e Intervenção Cognitivo-Comportamental (CINEICC), Coimbra, Portugal
5. Centro de Linguística da Universidade de Lisboa (CLUL), Lisbon, Portugal
6. Centro de Linguística da Universidade do Porto (CLUP), Porto, Portugal
7. Centro de Linguística da Universidade Nova de Lisboa (CLUNL), Lisbon, Portugal
8. Centro Algoritmi, Ponta Delgada, Portugal
9. Centro de Psicologia da Universidade do Porto (CPUP), Porto, Portugal
10. Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal
11. Universidade de Évora, Évora, Portugal
12. Instituto de Engenharia de Sistemas e Computadores (INESC), Lisbon, Portugal
13. INESC TEC, Laboratório de Inteligência Artificial e Apoio à Decisão (INESC TEC/LIAAD), Tomar
14. Instituto de Telecomunicações Coimbra (IT Coimbra), Coimbra, Portugal
15. Laboratório de Ciência da Computação e Informática (NOVA LINCS), Lisbon, Portugal
16. Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC), Porto, Portugal
17. Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brazil
18. Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil
19. Universidade de São Paulo, São Carlos, Brazil
20. Camões - Instituto da Cooperação e da Língua, Lisbon, Portugal

Core user communities

Data scientists, Software developers, Linguists and language scholars, NLP scholars, Healthcare software developers, Language teaching scholars, Computer scientists, Cognitive scientists, Digital humanities scholars, Language technology researchers.

Website

https://portulanclarin.net

Key achievements

Events
a. *Algorithms* [TV Debate] (2023, December 7) - part of the TV program "Civil Society" (in Portuguese), at the national public TV broadcasting RTP2 channel, about AI, societal impacts and sovereignty https://www.rtp.pt/play/p11183/e733134/sociedade-civil
b. *Artificial intelligence* [TV Debate] (2023, December 22) - at the MyGiG.pt channel https://youtu.be/pMFQ30RqhhE

Projects
a. *Covid-recovery program* funded project on generative AI for the Portuguese language

Resources (datasets, tools, services)
a. Collection of reference LLM encoders and decoders for Portuguese, open source, license and distribution, for both PT-PT and PT-BR variants, 100-900-1.5k Million parameters

Collaborations (notable collaborations with other academic institutions, and/or GLAM sector or industry)
a. New collaboration protocol with Lisbon Academy of Sciences

Publications
a. Rodrigues, J., Gomes, L., Silva, J., Branco, A., Santos, R., Lopes Cardoso, H., & Osório, T. (2023). Advancing neural encoding of Portuguese with transformer Albertina PT-*. In N. Moniz, Z. Vale, J. Cascalho, C. Silva, & R. Sebastião (Eds.), *Progress in Artificial Intelligence (EPIA 2023) Lecture Notes in Computer Science 14115*. https://doi.org/10.1007/978-3-031-49008-8_352
b. Branco, A., Eskevich, M., Frontini, F., Hajič, J., Hinrichs, E., de Jong, F., Kamocki, P., König, A., Lindén, K., Navaretta, C., Piasecki, M., Piperidis, S., Pitkänen, O., Simov, K., Skadiņa, I., Trippel, T., Witt, A., & Zinn, C. (2023). The CLARIN infrastructure as an interoperable language technology platform for SSH and beyond. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-023-09658-z
c. Branco, A. (2023). Línguas, Tecnologia e Inovação - Síntese. *CILPE2022: Línguas Cultura, Ciência e Inovação, OEI Organização do Estados Ibero-americanos*, 179-182. http://www.di.fc.ul.pt/~ahb/pubs/2023Branco.pdf

d. Branco, A., Grilo, S., & Silva, J. (2023). Language report on Portuguese. In G. Rehm & A. Way (Eds.), *European Language Equality*, Chapter 30. Springer. http://www.di.fc.ul.pt/~ahb/pubs/2023BrancoGriloSilva.pdf

e. Laborinho, A. P., & Branco, A. (2023). Portuguese language and technology: The future is now. http://www.di.fc.ul.pt/~ahb/pubs/2023LaborinhoBranco_EN.pdf [Originally published in Portuguese: (2023, February 9). Língua portuguesa e tecnologia: o futuro é agora. Público, 9. http://www.di.fc.ul.pt/~ahb/pubs/2023LaborinhoBranco.pdf]

Slovenia/CLARIN.SI

Coordinator name and institution

Tomaž Erjavec, Jožef Stefan Institute

List of consortium partners (with centre types hosted)

1. Jožef Stefan Institute (B-centre, K-centre)
2. Amebis, d.o.o.
3. Alpineon, d.o.o.
4. Institute of Contemporary History
5. Slovenian Language Technologies Society
6. National and University Library
7. University of Ljubljana
8. University of Maribor
9. University of Nova Gorica
10. University of Primorska
11. Scientific and Research Centre of the Slovenian Academy of Sciences and Arts
12. Science and Research Centre Koper

Core user communities

Corpus linguists, Language teaching scholars, Lexicographers, Computational linguists, Language technology researchers, Digital humanities scholars.

Website

https://www.clarin.si/

Key achievements

Events
a. *CLARIN.SI services* [Course] (February 2023) - Full day course for Slovenian librarians

Projects
a. Key partner in *ParlaMint II: Towards Comparable Parliamentary Corpora* (CLARIN ERIC)
b. *Development of Slovenian in a Digital Environment* - Ministry of Culture of Slovenia

Resources (datasets, tools, services)

a. Ljubešić, N., Kuzman, T., et al. (2024). CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation. P*roceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 3271–3282.

b. Erjavec, T., et al. (2024). *ParlaMint 4.1: Multilingual comparable corpora of parliamentary debates*. [with 29 parliamentary corpora and over 1 billion words, also machine translated into English].

c. *CLASSLA* - online annotation tool for South-Slavic languages https://www.clarin.si/info/k-centre/comparable-classla-web-corpora-of-south-slavic-languages/

d. *SENTA* - online tool for sentence simplification and analysis https://senta.cjvt.si/

e. *Korpusnik* - online tool for quick overviews of word usage https://korpusnik.cjvt.si/en

Collaborations (notable collaborations with other academic institutions, and/or GLAM sector or industry)

a. The National and University Library of Slovenia joined the CLARIN.SI consortium in 2023

Publications

a. Ljubešić, N., & Samardžić, T. (2023). *Together we are stronger: Collaborative development of language resources and technologies for South Slavic languages* [Lecture]. 9th edition of the Slavic NLP Workshop. https://www.clarin.si/info/wp-content/uploads/2023/05/Together-we-are-stronger_-Collaborative-development-of-resources-and-technologies-for-South-Slavic-languages.pdf

b. Ljubešić, N. (2023, August). *Building a linguistic corpus infrastructure: The South-Slavic example* [Lecture]. Conference Corpora in Language learning, Translation and Research, University of Zadar. https://www.clarin.si/info/wp-content/uploads/2023/09/Building-a-linguistic-corpus-infrastructure_-The-South-Slavic-example.pdf

c. CLARIN.SI: The Slovenian node of CLARIN [Poster]. (2023, July 10-14). *Digital Humanities Conferenc*e. Graz, Austria.

Other

a. Prepared and published the CLARIN.SI strategy 2024-2030.

South Africa/SADiLaR	CLARIN-ZA

Coordinator name and institution

Friedel Wolff (SADiLaR, North-West University)

List of consortium partners (with centre types hosted)

Only SADiLaR is a CLARIN centre (currently C-centre). Our partners are the SADiLaR nodes:

1. University of Pretoria (Department of African Languages)
2. University of South Africa (Department of African Languages)
3. CSIR (HLT Research Group)

4. North-West University (Centre for Text Technology)
5. Inter-Institutional Centre for Language Development and Assessment (ICELDA)
6. Child Language Development Node (Department of General Linguistics at Stellenbosch University)

Core user communities

Digital humanities scholars, Language technology researchers, Linguists and language scholars, Social science scholars, Computer scientists, Linguists and language scholars, Literary scholars, Speech technology researchers, Social science scholars.

Website

Website: https://sadilar.org/
ESCALATOR Community

Key achievements

Events
a. *2 DH IGNITE Events* - Regional events that aim to grow the community and build capacity for the humanities and social sciences across South African Universities
   - *DH IGNITE: Western Cape* (2023, March 8-10) - https://sadilar.org/en/dh-ignite-western-cape/
   - *DH IGNITE: Eastern Region* (2023, August 23-25) - https://sadilar.org/en/dh-ignite-digital-upskilling-in-the-spotlight-at-third-regional-event/
b. *10 Digital Humanities Colloquiums* - Online DH colloquiums as part of SADiLaR's ESCALATOR programme. These colloquiums cover a wide variety of topics related to

Digital Humanities, focusing on techniques or methodologies used, as well as applications or application areas.

c. *Supported Conferences* - SADiLaR provided support and played a central role in organising national and regional conferences:

- *The South African Linguistics and Applied Linguistics Society (SALALS)* (2023) - https://sadilar.org/en/salals-conference/
- *The African Association for Lexicography (AFRILEX)* (2023) - https://sadilar.org/en/sadilar-shines-spotlight-on-research-during-spring-conference-season/
- *The African Language Association of Southern Africa (ALASA)* (2023) - https://alasa.org.za/conferences/
- *The Digital Humanities Association of Southern Africa (DHASA)* (2023) - https://www.bizcommunity.com/article/dhasa-2023-highlights-importance-of-digital-humanities-for-inclusion-123006a?fbclid=IwZXh0bgNhZW0CMTAAAR3Ixqh5BY_2gI-OIjRfjVesk8XMCaNKY-1itEpq1NE2avQrWPRcLAVqST4_aem_AURrbVRfwPXK2eAaK6YGgm_ILskvswkXqKn6iTYbt6mQvHq0tTxbLPRqLeJ4NBkMJT12AIDi7bqSgCYZbJ_AYwxp
- *The International Language Conference* (2023) - hosted at Sol Plaatje University https://www.spu.ac.za/index.php/spu-hosts-inaugural-international-language-conference/
- *The Resources for African Indigenous Languages (RAIL)* (2023) - https://sadilar.org/en/sadilar-celebrates-successful-4th-rail-workshop-in-croatia/

Projects

a. *SWiP project launch* - SADiLaR launched its SWiP project in collaboration with Wikipedia and the Pan-South African Language Board. The launch event focused on the Ndebele language, however, the project promotes all of South Africa's indigenous languages online: https://sadilar.org/en/swip-project-to-champion-sas-indigenous-languages-online/

b. *SADiLaR* - concluded the USAf Language Resources Audit that investigated the resources available and still required for institutions to implement the policy framework for public higher education institutions: https://sadilar.org/index.php/en/language-resources-audit

Resources (datasets, tools, services)

a. The SADiLaR repository received technical upgrades during 2023 and prepared the work for future improvements. New resources in the repository in 2023 include:

- *80 datasets* with word embedding models for 10 South African languages https://repo.sadilar.org/discover?filtertype_1=project&filter_relational_operator_1=equals&filter_1=NCHLT+Text+IV&filtertype_2=title&filter_relational_operator_2=contains&filter_2=&submit_apply_filter=&query=embeddings
- *Afrikaans lexical blends dataset* https://repo.sadilar.org/handle/20.500.12185/668

Collaborations (notable collaborations with other academic institutions, and/or GLAM sector or industry)

The USAf audit (mentioned above) was in collaboration with all South African public universities, and many of our other projects and events (see above) are hosted in collaboration with another institution.

Book chapters:

Van der Walt, A., Steyn, J., Trusler, A., & Van Zaanen, M. (2023). Challenges and opportunities of Digital Humanities training in South Africa: Moving beyond the silos. In L. Estill & J. Guiliano (Eds.). *Digital Humanities Workshops: Lessons Learned (1st ed.)*. Routledge. https://doi.org/10.4324/9781003301097

Journal articles:

Trollip, B. (2023). 'n Gebruiksgebaseerde beskrywing van Afrikaanse prefiksoïede. *LitNet Akademies 20*(3). https://doi.org/10.56273/1995-5928/2023/j20n3g1

Mahlaza, Z., & Khumalo, L. (2023). Algorithm for assisting grammarians when extracting phonological conditioning rules for Nguni languages. Journal of the Digital Humanities Association of Southern Africa 5(1).

Keet, C. M., & Khumalo, L. (2023). Contextualizing levels of language resourcedness affecting digital processing of content. *Digital Humanities Quarterly (DHQ): The Alliance of Digital Humanities Organizations*.

Publications

a. **Reports:**
   - Van Dyk, T.J., Wilken, I., Ravyse, N., Du Plessis, D., Khumalo, L., & Steyn, J. (2023). National Language Resources Audit. repo.sadilar.org/handle/20.500.12185/667
b. **Blog posts:**
   - du Plessis, D. (2023, February 21). *Exploring digital humanities in South Africa: A discussion with Justus Roux* [Blog post]. https://dh-ignite.org/post/2023/02/conversation-justus-roux/
   - du Plessis, D. (2023, January 30). *Bridging the Gap: Connecting South African Humanities and Social Sciences through DH-IGNITE* [Blog post]. https://dh-ignite.org/post/2023/01/building-bridges/
c. **Conferences / Oral presentations:**
   - Mlambo, R., Matfunjwa, M., & Skosana, N. (2023). A Morphological- and Semantic Analysis of Compound Words in Xitsonga. *Southern African Linguistics and Applied Linguistics Society Conference 2023*. Walter Sisulu University.
   - Matfunjwa, M., Mlambo, R., & Skosana, N. (2023). Nicknames Among Swati Clans: A Socio-Cultural Analysis. *Southern African Linguistics and Applied Linguistics Society Conference 2023*. Walter Sisulu University.
   - Skosana, N., Matfunjwa, M., & Mlambo, R. (2023). *Reduplication of Nouns and Verbs in isiNdebele. Sol Plaatje International Language Conference Kimberley*.
   - Nomsa S., Mlambo R., & Matfunjwa, M. (2023). *Strategies for creating corpora and language resources for under-re sourced South African indigenous languages*. 10th

Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (21-23 April 2023), Poznań, Poland.

Training materials
a. First steps to R for Humanities and Social Sciences
b. First steps to Python for Humanities and Social Sciences

Spain/CLARIAH-ES

CLARIAH-ES

Coordinator name and institution

German Rigau (HiTZ Basque Center of Language Technology, UPV/EHU)

List of consortium partners (with centre types hosted)

1. UPV/EHU (HiTZ)
2. Universidad de Santiago de Compostela (Instituto da Lingua Galega and CiTIUS)
3. Universidad de Alicante (Biblioteca Virtual Cervantes)
4. UNED (LINDH and LENAR)
5. Barcelona Supercomputing Center- Centro Nacional de Supercomputación (BSC-CNS)
6. Universidad Complutense de Madrid (UCM)
7. Universidad de Jaén (CEATIC)
8. Universidad de las Palmas de Gran Canaria (IATEXT)
9. CSIC (Centro de Ciencias Humanas y Sociales)
10. Biblioteca Nacional de España (BNE)

Core user communities

Digital humanities scholars, Linguists and language scholars, Social science scholars, Computational linguists.

Website

http://clariah.es
http://clariah.eus
https://www.ucm.es/humanidadesdigitales/centro-clariahes
https://clariah.ujaen.es/

Key achievements

Events
a. *Workshop CLARIAH-CSIC* (2023, May 11) - Madrid
   https://digital.csic.es/handle/10261/310886
b. *Kick-off CLARIAH-ES* (2023, June 15-16) - Alicante https://ixa2.si.ehu.eus/red-clariah-es/actividades
c. *Bibliotecas, datos, inteligencia artificial, BNE* (2023, November 7-8) - Madrid
   https://www.bne.es/es/agenda/bibliotecas-datos-inteligencia-artificial-nuevas-rutas-conocimiento

   d. *Workshop CLARIAH-EUS* (2023, November 23) - Donostia
      https://www.donostiakultura.eus/es/cursos/clariah-eus-euskararako-ikerketa-azpiegitura-eraikitzen-cas
   e. *CLARIAH-ES presentation at UJA* (2023, December 14) - Jaén
      https://clariah.ujaen.es/2023/11/23/1era-jornada-uja/

Projects

   a. *Red Clariah-es* - Red estratégica para la integración en las infraestructuras europeas de investigación en Ciencias Sociales y Humanidades https://ixa2.si.ehu.eus/red-clariah-es/
   b. *CLARIAH-EUS* - Europako ikerketa-azpiegituretan Giza eta Gizarte Zientzietan euskara eta euskaraz ikertzeko aukera bultzatzeko egitasmoa. https://www.clariah.eus/en
   c. *ILENIA Project* - Boosting Languages in Artificial Intelligence. https://proyectoilenia.es/en/
   d. *Toponomasticon Hispaniae* - https://toponhisp.org/
   e. *CLS INFRA* - https://clsinfra.io/

Resources (datasets, tools, services)

   a. *Analhitza* - https://ixa2.si.ehu.eus/clarink/analhitza.php
   b. Arquivo do Galego Oral (AGO) - https://ilg.usc.gal/ago/
   c. *UVIGO-TALG CLARIN K-centre: DContado* - https://ilg.usc.gal/dcontado/index.php?lang=en
   d. *Estilometría TIP* - https://tulengua.iatext.ulpgc.es/estilometria/

Publications

   a. Agerri, R., Agirre, E., Aldabe, I., Aranberri, N., Arriola, J. M., Atutxa, A., Azkune, G., Campos, J. A., Casillas, A., Estarrona, A., Farwell, A., Goenaga, I., Goikoetxea, J., Gojenola, K., Hernáez, I., Iruskieta, M., Labaka, G., Lopez de Lacalle, O., Navas, E., Oronoz, M., Otegi, A., Pérez, A., Perez de Viñaspre, O., Rigau, G., Salaberria, A., Sanchez, J., Saratxaga, I., & Soroa, A. (2023). State-of-the-art in language technology and language-centric artificial intelligence. In G. Rehm, & A. Way (Eds.). *European Language Equality: Cognitive Technologies.* Springer. https://doi.org/10.1007/978-3-031-28819-7_2
   b. Boullón Agrelo, A. I. (2023). Toponomasticon Hispaniae: planteamiento, objetivos y recursos bibliográficos (con especial atención a Galicia y Portugal). In J. Terrado Pablo & J. Giralt Latorre (Eds.), *Nuevos senderos en la toponimia hispánica* (pp. 15-40). Pagès Editors.
   c. Domínguez Vázquez, M. J., & Gouws, R. H. (2023). The Definition, Presentation and Automatic Generation of Contextual Data in Lexicography. *International Journal of Lexicography 36*(3), 233-259. https://academic.oup.com/ijl/article/36/3/233/7153265
   d. Erjavec, T., et al. (2023). *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint. 4.0*. http://hdl.handle.net/11356/1860.
   e. Baucells de la Peña, I., Calvo Figueras, B., Villegas, M., & Lopez de Lacalle, O. (2023). Entailment-based Task Transfer for Catalan Text Classification in Small Data Regimes *Procesamiento del Lenguaje Natural 71*(Sep), 165-177. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6551

Sweden/SWE-CLARIN

SWE-CLARIN

Coordinator name and institution

Lars Borin (Språkbanken Text, University of Gothenburg)

List of consortium partners (with centre types hosted)

1. Språkbanken Text, University of Gothenburg (B)
2. Språkbanken Tal, Royal Institute of Technology (K)
3. Språkbanken Sam, Institute for Language and Folklore (K)
4. Gothenburg Research Infrastructure in Digital Humanities, University of Gothenburg (K)
5. The National Library of Sweden
6. The Swedish National Archives (K)
7. Linköping University (K)
8. The Humanities Lab, Lund University (C, K)
9. Stockholm University (K)
10. Umeå University
11. Uppsala University (K)

Core user communities

Linguists and language scholars, Digital humanities scholars, Literary scholars, Conceptual and modern historians, Historians, Ethnologists, Social science scholars, Political science scholars, Sociologists, Archive scholars.

Website

https://sweclarin.se/eng/home

Key achievements

Events

a. *Swe-CLARIN on tour* (2023, February 1) - Uppsala University, Topic: Researching historical archival texts with AI and language technology
b. *Seminar on language technology for the national minority language meänkieli* (2023, February 27)
c. *Methods for handling of text data* [PhD Course] (7.5 ECTS HEC) (Spring 2023) - Stockholm University https://www.su.se/doctoral-school-in-the-humanities/courses/methods-for-handling-of-text-data-7-5-ects-1.633720
d. *Språkbanken's autumn workshop 2023* (2023, October 23) - Workshop topic: research on linguistic variation

Projects
  a. *HUMINFRA* - https://www.huminfra.se/
  b. *Grandma Karl is 27 years old* - Automatic pseudonymization of research data
     https://spraakbanken.gu.se/en/projects/mormor-karl

Resources (datasets, tools, services)
  a. *SuperLim 2* - Swedish counterpart of the SuperGLUE NLU benchmark for English
     https://spraakbanken.gu.se/en/resources/superlim
  b. *FOLKE* - Web interface to the vast archive of digitized historical texts of the Institute for
     Languages and Folklore https://forska.folke.isof.se/#/search/analyse
  c. *Visible Speech (VISP)* - https://www.umu.se/en/research/infrastructure/visible-speech/
  d. *Mink* - Språkbanken's data platform https://spraakbanken.gu.se/mink/
  e. *Swedish Sign Language lexicon* - https://teckensprakslexikon.su.se/
  f. *Swedish Sign Language corpus (with UD annotation)* -
     https://universaldependencies.org/swl/

Collaborations (notable collaborations with other academic institutions, and/or GLAM sector or industry)
  a. *DigPhil* (2023–2027) - National PhD school in digital philology
     https://www.lingfil.uu.se/forskning/digphil

Publications
  a. Berdicevskis, A., Bouma, G., Kurtz, R., Morger, F., Öhman, J., Adesam, Y., Borin, L.,
     Dannélls, D., Forsberg, M., Isbister, T., Lindahl, A., Malmsten, M., Rekathati, F., Sahlgren,
     M., Volodina, E., Börjeson, L., Hengchen, S., & Tahmasebi, N. (2023). Superlim: A Swedish
     Language Understanding Evaluation Benchmark. *Proceedings of the 2023 Conference on
     Empirical Methods in Natural Language Processing,* 8137–8153. Association for
     Computational Linguistiscs. https://doi.org/10.18653/v1/2023.emnlp-main.506
  b. Borin, L., Domeij, R., Edlund, J., & Forsberg, M. (2023). Language Report Swedish. In G.
     Rehm & A. Way (Eds.), *European Language Equality*, 219–222. Springer.
  c. Fallgren, P., & Edlund, J. (2023). Crowdsource-based validation of the audio cocktail as a
     sound browsing tool. *Proceedings of INTERSPEECH 2023*, 2178–2182.
  d. Fridlund, M., Azar, M., Brodén, D., & McGuire, M. (2023). The Cultural Imaginary of
     Terrorism: Close and Distant Readings of Political Terror in Swedish News and Fiction
     During the Cold War. *Digital Humanities in the Nordic and Baltic Countries Publications* 5
     (1), 90–104.
  e. Hammarlin, M.M., Kokkinakis, D., & Borin, L. (2023). COVID-19 Vaccine Hesitancy: A Mixed
     Methods Investigation of Matters of Life and Death. *Journal of Digital Social Research
     (JDSR)* 5 (4), 31–61.
  f. Volodina, E., Mohammed, Y. A., Berdicevskis, A., Bouma, G., & Öhman, J. (2023). DaLAJ-
     GED - a dataset for Grammatical Error Detection tasks on Swedish. *Proceedings of the
     12th Workshop on Natural Language Processing for Computer Assisted Language Learning
     (NLP4CALL 2023)*, 94–101. LiU EP. https://doi.org/10.3384/ecp197011

Switzerland/CLARIN-CH	CLARIN-CH

Coordinator name and institution

Cristina Grisot (University of Zurich)

List of consortium partners (with centre types hosted)

1. University of Basel
2. University of Bern
3. University of Geneva
4. University of Fribourg
5. University of Lausanne
6. University of Neuchâtel
7. University of Zurich
8. Università della Svizzera italiana
9. Zurich University of Applied Sciences
10. 10. Swiss Academy for Human and Social Sciences

Core user communities

Linguists and language scholars, NLP scholars, Computational linguists, discourse scholars, Pragmatics scholars, Language teaching scholars, Data scientists, Digital humanities scholars, Psycholinguists, Neurolinguists, Speech technology researchers.

Website

https://clarin-ch.ch (main website, where all our activities are shown)
https://www.liri.uzh.ch/en.html (service provider)
https://www.lars.uzh.ch/en.html (repository and data service unit)
https://www.zhaw.ch/en/linguistics/research/swiss-al/ (web platform with language data)

Key achievements

Events

a. *UPSKILLS: Linguists in tech: closing the skills gap* [Workshop] (2023, April 21-22) - https://clarin-ch.ch/news/2023/02/21_1686023495
b. *Navigating Legal Complexities: Open Research Data & Language Corpora* [Round Table] (2023, November 28) - https://clarin-ch.ch/news/2023/10/17_1697525420
c. *CHORD-Talk-in-interaction project* [Kick-off event] (2023, April 28) - https://clarin-ch.ch/news/2023/03/29_1686023715
d. *Management of sensitive and personal data: Ethical & legal aspects for linguistic data in Switzerland* [Kick-off event] (2023, September 29) - https://clarin-ch.ch/news/2023/06/15_1686822091

e. Designing, building and using data banks of interactional corpora from a conversation analytic perspective [Workshop] (2023, December 7-8) - https://clarin-ch.ch/news/2023/10/31_1698758927

Projects
a. *Upgrading the linguistic ORD-ecosystem: UpLORD* - project funded thanks to the national Open Research Data programme https://www.liri.uzh.ch/en/projects/UpLORD.html
b. *KIParla corpus* - project funded thanks to the national Open Research Data programme https://www.chord-talk-in-interaction.usi.ch/cti/workshops/workshop-1
c. *Swiss-AL: Linguistic Open Research Data Practices for Applied Sciences* - project funded thanks to the national Open Research Data programme https://www.zhaw.ch/en/linguistics/research/swiss-al-linguistic-open-research-data-practices-for-applied-sciences/
d. *Strengthening ORD competences in the Applied Sciences: Open Educational Resources for Swiss-AL* - project funded thanks to the national Open Research Data programme https://www.zhaw.ch/en/research/research-database/project-detailview/projektid/6990/

Resources (datasets, tools, services)
a. *Digital Linguistics Workbench* - https://swiss-al.linguistik.zhaw.ch/rsconnect/workbench
b. *LiRI Corpus Platform (LCP)* - https://liri.linguistik.uzh.ch/wiki/langtech/lcp/start

Publications
a. Vamvas, J., Graën, J., & Sennrich, R. (2023). SwissBERT: The Multilingual Language Model for Switzerland. *8th Swiss Text Analytics Conference (SwissText), Neuchâtel, Switzerland, 12 June 2023 - 14 June 2023*, 54-69. Association for Computational Linguistics.
b. Krasselt, J., Dreesen, P., Stücheli-Herlach, P., Lemmenmeier, D., Cho, S., Rothenhäusler, K., & Fluor, M. (2023). Swiss-AL: platform for language data in applied sciences: on challenges in the field of language open research data. *1st Conference on Research Data Infrastructure (CoRDI)-Connecting Communities, Karlsruhe, Germany, 12-14 September 2023*, 1. TIB Open Publishing. https://doi.org/10.21256/zhaw-29441
c. Schaber, J., Graën, J., McDonald,D., Mustac, I., Rajovic, N., Schneider, G., Bubenhofer, N., Linden, K., Niemi, J., and Kontino, T. (2023). The LiRI Corpus Platform. *CLARIN annual conference proceedings*, 145-149. CLARIN ERIC.
d. Bubenhofer, N., Malits, A., Strebel, S., Gräen, J., Buerli, S., & Grisot, C. (2023). Building and consolidating a FAIR-compliant ecosystem of infrastructures. *CLARIN Annual Conference Proceedings*, 95.
e. Graën, J.,, Mustac, I., Rajovic, N., Schaber, J., Schneider, G., Bubenhofer, N., Linden, K., Niemi, J., & Kontino, T. (2023). Swissdox@ LiRI–a large database of media articles made accessible to researchers. *CLARIN annual conference proceedings*, 111-115. CLARIN ERIC.

Other
a. *CLARIN-CH Documentation Platform* - https://clarin-ch.ch/documentation-platform/start
b. *Social interaction is among people: Legal, technical, and ethical explorations about personal information and its removal in talk-in-interaction as data* [Documentation] - https://www.chord-talk-in-interaction.usi.ch/news/feeds/36387

United Kingdom/CLARIN-UK	

Coordinator name and institution

Martin Wynne (University of Oxford)

List of consortium partners (with centre types hosted)

1. British Library
2. Centre for Corpus Research, University of Birmingham
3. Corpws Cenedlaethol Cymraeg Cyfoes / The National Corpus of Contemporary Welsh, Cardiff University
4. School of Humanities, Coventry University
5. Language Technology Group, University of Edinburgh
6. School of Critical Studies, University of Glasgow
7. King's Digital Laboratory, King's College London
8. CASS and UCREL, Lancaster University
9. Centre for Translation Studies, University of Leeds
10. Faculty of Linguistics, Philology and Phonetics, University of Oxford (Oxford Text Archive, C-centre)
11. Natural Language Processing Group, University of Sheffield
12. Centre for Transnational and Transcultural Research (CTTR), University of Wolverhampton

Core user communities

Linguists and language scholars, Literary scholars, Historians, Computer scientists, NLP scholars.

Website

https://www.clarin.ac.uk/
https://llds.ling-phil.ox.ac.uk/ (Repository and C-centre)

Key achievements

Events

a. *Lancaster Talks on Language: Launch of #LancsBox X application* [online seminar] (2023, February 24) - in collaboration with CLARIN-UK, attended by 1300 people
b. *Making Holocaust Oral Testimonies More Usable as Research Data* (2023, May 15-17) - joint CLARIN-EHRI research workshop organized by CLARIN-UK and EHRI-UK, King's College, London

   c. *What can you do with the CLARIN research infrastructure?* [pre-conference workshop] (2023, July 2) - Corpus Linguistics 2023 Conference, Lancaster University
   d. *CLARIN Café: A new CLARIN Resource Family for Lexical Semantic Change Research* [online webinar] - organised by Barbara McGillivray (King's College London)
   e. *CLARIN and Libraries 2023: Large Language Models and Libraries* (2023, December 5-6) - co-organized with CLARINO, National Library of Norway, Oslo

Projects
   a. *The Sheffield NLPG* joined the EU project *ATRIUM* as a partner thanks to their status as a member of the CLARIN-UK consortium, and are providing NLP expertise and services to this large-scale archaeology project
   b. *CLARIN Resource Families project on lexical semantic change* (Nov 2022- June 2023) - led by Barbara McGillivray, King's College London. Funded by CLARIN ERIC

Resources (datasets, tools, services)

New datasets in the CLARIN-UK repository:
   a. Wynne, M. (2024). *Frequently repeated clusters of words in Early English Books Online*. Literary and Linguistic Data Service. http://hdl.handle.net/20.500.14106/2570
   b. McGillivray, B. (2021). *Latin lexical semantic annotation*. Literary and Linguistic Data Service. http://hdl.handle.net/20.500.14106/2568
   c. Lähteenoja, V. & Vatri, A. (2021). *Ancient Greek semantic annotation datasets*. Literary and Linguistic Data Service. http://hdl.handle.net/20.500.14106/2569
   d. Gwillim, E., & Symonds, M. (2023). *The Letters of Elizabeth Gwillim and Mary Symonds, 1801-1808*. Literary and Linguistic Data Service. http://hdl.handle.net/20.500.14106/2567
   e. Knight, D. (2020). *CorCenCC: Corpws Cenedlaethol Cymraeg Cyfoes: The national corpus of contemporary Welsh*. Literary and Linguistic Data Service. http://hdl.handle.net/20.500.14106/2564
   f. Erjavec, T., et al. (2023). *Multilingual comparable corpora of parliamentary debates ParlaMint 4.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1859 [UK parliamentary proceedings and multilingual annotations contributed to the ParlaMint corpus by Lancaster University]

Collaborations (notable collaborations with other academic institutions, and/or GLAM sector or industry)
   a. CLARIN-UK has led a notable collaboration on behalf of CLARIN with libraries. The collaboration has resulted in two workshops, two conference presentations and a presentation to the DARIAH national coordinators' meeting, and the setting up of a The collaboration provides important links between major projects developing online interfaces, data labs and large language models, and is making an important contribution to sharing expertise and knowledge, and promoting interoperability. The activities have involved national, university and other research libraries from sixteen countries.
   b. CLARIN-UK has also initiated and led a notable collaboration with the European Holocaust Research Infrastructure (EHRI), with a presentation of CLARIN at an EHRI-UK consortium meeting, a joint research workshop, and successful proposals for further workshops in 2024, with research grant proposals also planned.

c. The Sheffield centre participates in the ATRIUM-EU project as CLARIN representatives with partners in archaeology and digital humanities.
d. The Infrastructure in the DIgital Arts and Humanities (iDAH) programme provides funding for the CLARIN-UK repository, and also acts as a forum for peer trusted repositories the Archaeology Data Service, Natural History Museum and British Library to collaborate on common practices, training and user services.

Training materials
a. *CLARIN Resource Families project on 'Desktop and Online Tools for Corpus Analysis'* - completed and published online by CLARIN (https://www.clarin.eu/resource-families/corpus-query-tools)
b. *How to Search for Linguistic Patterns across Corpora* - https://www.clarin.eu/sites/default/files/CL2023_Tutorial Handout.pdf

Other
a. Paul Rayson (Lancaster) served an additional year as CLARIN Ambassador, contributing to a number of the events and collaborations above.