CLARIN AGREEMENT

CLARIN Agreement with the National Consortium of Slovenia

(cf. Annex 1-A for the composition of the National Consortium)

This Agreement is made by and between:

(1) CLARIN ERIC, a European Research Infrastructure Consortium established by Commission Implementing Decision (EU) 2012/136 of 29 February 2012 (hereinafter "**CLARIN ERIC**"), and,

(2) the National Consortium of Slovenia, represented by Jožef Stefan Institute (ID SI55560822) (hereinafter "**National Consortium**").

Hereinafter collectively referred to as the "Parties" and individually as "Party".

HAVE AGREED on the following specific terms and conditions for contribution of resources to CLARIN ERIC during the period

1.1.2024 – 31.12.2027

Article 1 – Scope

The Parties have agreed to conclude an agreement in accordance with Article 6.5 of the Statutes, including the specific terms and conditions described in this Agreement which relate to Art. 2, 6.1 and 6.2 (d)-(m) of the Statutes.

Article 2 – Definitions

"National Coordinator" means the person appointed as a national coordinator by a CLARIN ERIC member in accordance with Article 6(2)(e) of the Statutes and which is responsible for the National Consortium.

"Statutes" means the Statutes of CLARIN ERIC including its Annexes.

Article 3 - Obligations

3.1 Contribution to CLARIN ERIC overall and specific obligations listed in articles 2 and 6 of the statutes

In addition to the specific contributions mentioned below in Article 3.2 below, the National Consortium shall (i) promote the adoption of relevant standards in national resource and tool creation projects, as well as the uptake of CLARIN services among researchers in their country, (ii) gather user feedback and requirements, and (iii) support CLARIN centres in the member country by facilitating integration into national and other relevant infrastructures.

3.2 Specific National Contributions

This section describes the national level in-kind contribution to CLARIN; part of this contribution will be coordinated at the transnational level through CLARIN ERIC, and part will be coordinated at the national level. This agreement lists the components of the offering at the time of signature. For each component, details are provided in technical annexes where relevant.

3.2.1. Contributions to be coordinated by CLARIN ERIC

(a) **Data and service centres**: In accordance with Art. 6.2 (f) of the Statutes, the National Consortium shall contribute at least one centre of type B (cf. the check list https://office.clarin.eu/v/CE-2013-0095-B-centre-checklist-v6.pdf).

Detailed descriptions of the centres and the services they offer are included in Annex 1-B. CLARIN centres are expected to obtain a deep level of integration with the infrastructure by connecting a substantial portion of their language resources and tools via:

- harvestable metadata for the VLO
- endpoints for the Federated Content Search
- web applications for the LR Switchboard

(b) **Access**: The National Consortium shall comply with Article 19 of the Statutes on access policies for users. Any restrictions or exceptions are specified in Annex 1-F, to the extent it is applicable.

(c) **User Authentication and Authorization System**: The National Consortium shall comply with the specification given in Annex 1-C.

(d) **Knowledge sharing infrastructure (KSI)**: The National Consortium shall contribute to the KSI by providing access to its knowledge and expertise to the CLARIN community at large in accordance with Annex 1-D.

(e) **Membership of CLARIN ERIC committees and any other contribution** being offered to CLARIN and lending itself for coordination by CLARIN ERIC at the European level. Detailed descriptions are included in Annex 1-E.

3.2.2. Contributions coordinated at the national level, but compliant with CLARIN

The National Consortium shall provide for:
(a) Enhancements and improvements to existing resources, tools and services in accordance with the detailed descriptions included in Annex 2-A.

(b) Creation of new resources, tools and services in accordance with the detailed descriptions included in Annex 2-B.

(c) Measures to promote the adoption of and adherence to CLARIN standards in national resources and tools creation programmes and projects in accordance with the detailed descriptions included in Annex 2-C.

(d) Measures to create a knowledge-sharing infrastructure at the national level, with a view to promoting uptake of CLARIN services by the target audience, to exchanging knowledge and expertise, and to advancing Humanities and Social Sciences research through the support by CLARIN. A brief description of those items that are not already covered by 3.2.1-(d) is provided in Annex 2-D.

(e) Any other contribution relevant to the goals of CLARIN, including collaboration with third parties that CLARIN as a whole could contribute to or from which CLARIN could benefit. Detailed descriptions are included in Annex 2-E (if applicable).

3.3. National contributions made jointly with other parties

In accordance with Article 6.3 of the Statutes, members may decide to provide some of the national contributions in cooperation with other members, observers or third parties. In such cases, a separate multi-party agreement may be made between the CLARIN ERIC and those members / observers / third parties concerned.

Signatures

Date: 19 December 2023

Name: Tomaž Erjavec
Function: National Coordinator

Date: 20 December 2023

Name: Darja Fišer
Function: Executive Director, CLARIN ERIC

Annex 1 National contributions to be coordinated by CLARIN ERIC

Annex 1-A

The National Consortium of Slovenia, CLARIN.SI, consists of the following partner(s):

1. Alpineon, d.o.o.

Alpineon is a Slovenian RTD-performing SME specializing in developing state-of-the-art computer vision and speech-technology products. Alpineon's RTD team has extensive experience in hardware and software development, including CTI applications, VoIP devices and services, biometric technologies (speaker and face recognition), image processing (3-D vision) and speech technologies: Slovenian text-to-speech synthesis (TTS), automatic speech recognition (ASR), speech-to-speech translation (STS), voice portal applications etc.

Alpineon's RTD team is involved in several international and national research projects in the fields of language technologies, image processing and biometrics. It consists of 14 researchers and developers including 7 PhD holders. Alpineon is the recipient of the Slovenian Award for Technical Innovations for Disabled Persons in 2003, the award for Outstanding Research Achievements by the Slovenian Research Agency in 2013, the winner of the international ICB 2013 face recognition challenge and the ICB 2013 speaker recognition challenge and the recipient of the Slovenian Ambassador of Privacy award in 2014 for best practices in privacy protection.

Alpineon has been a member of CLARIN since 2007, and a founding member of CLARIN.SI. Alpineon contributes to CLARIN with language resources and speech technology engines.

2. Amebis, d.o.o.

The Amebis Company was established in 1991 for software development and production in the fields of language technologies and electronic publishing. The primary objective is to create core technologies (modules) and products for general use. The main areas of development and products are:

- Corpora: development, creation and management of text and speech corpora including more than 10 largest Slovenian corpora, with the 1 billion word Gigafida as the largest.
- Language processing: various language processing modules for Slovenian and some other languages, which are available as a plug-ins for various program packages (MS Office, Lotus Notes, SAS etc.): spell-checkers, hyphenators, lemmatizers, word form generators, grammar checkers.
- Machine translation systems: developers of Presis, a rule-based translation system, for Slovenian, which is part of the iTranslate4 translation system.
- Speech synthesis: the Govorec speech synthesizer (developed together with JSI) and high quality speech synthesizer eBralec (with Alpineon and JSI) for Slovenian language.
- Dialogue systems: SecondEgo is a platform for creating and dealing with virtual agents which helps websites and applications to simplify communication with the end users in different natural languages.
- Electronic dictionaries: preparation of more than 160 electronic and 80 paper dictionaries. The best known are portals Termania and Fran.

Amebis was also a partner in several EU projects.

Amebis was one of the founding members of CLARIN.SI and contributes to CLARIN primarily through language resources and language technology applications.

3. Jožef Stefan Institute

The Jožef Stefan Institute (JSI) is the leading Slovenian research organization for basic and applied research in natural sciences and technology with over 900 employees. Three organisational units are involved in developing and maintaining the CLARIN.SI infrastructure:

- The Department of Knowledge Technologies performs research in advanced information technologies, aimed at acquiring, storing and managing knowledge to be used in the development of knowledge-based applications. Established areas include intelligent data analysis, text and web mining, language technologies and computational linguistics, decision support and knowledge management. In the area of language technologies, the Department is one of the leading (and oldest) Slovenian centres for the development of language resources and annotation tools, esp. in the area of standardisation of resource encoding and linguistic formalisms and in open accessibility of resources. The Department was also among the first to develop and promote the area of Digital Humanities in Slovenia.
- The Artificial Intelligence Laboratory is concerned with research and development in information technologies with an emphasis on artificial intelligence. The main research areas are data analysis with an emphasis on text, web and cross-modal data; scalable real-time data analysis; visualization of complex data; semantic technologies; and language technologies. The Laboratory has been involved in many EU projects in the area of text analytics and processing, where their task is mainly in providing technologies for knowledge extraction from text and for machine translation. The Laboratory also puts special emphasis on the promotion of science. In collaboration with the Centre for Knowledge Transfer in Information Technologies (CT3) they are developing the award-winning VideoLectures.NET educational portal and organizing the national ACM competition in Computer Science (in Slovene).
- The Networking Infrastructure Centre manages the networking and hardware infrastructure at JSI. It is active on the areas of trust and authentication, e.g. it maintains the JSI IdP service and is actively involved in EduGain and other EU efforts in these areas.

The JSI is the host of the CLARIN.SI research infrastructure. It coordinates the work of the infrastructure, maintains and develops its repository and services, provides language resources and tools etc.

4. Institute of Contemporary History

The Institute of Contemporary History (ICH) is the central national institution for historiographical research of the period form 19th century to today. The Institute is one of the most important institutions in the Digital Humanities in Slovenia, and is the national coordinating institution for DARIAH-SL, a member of DARIAH ERIC.

One of the three research programmes of the Institute is the Research Infrastructure of the Slovenian Historiography, which is maintaining the SIstory Web portal of Slovenian Historiography. The RI performs digitisation of material of historic importance and management of digitally born content. Another task is online publishing of basic sources for historiographical research and literature from the field, regardless of its form or format, thus not limited only to textual sources. Since textual files represent the majority of its digital archival holdings, the focus is given to advanced text mark-up, using the TEI Guidelines and relevant XML technologies. Large textual corpora are being created following this process, based mainly on the stenographical minutes of different Slovene legislative bodies

ICH was one of the founding members of CLARIN.SI and contributes to CLARIN primarily through making available its large and richly encoded text collections of recent Slovenian historical sources and as being the primary liaison to DARIAH(-SI).

5. National and University Library of Slovenia
The National and University Library (NUK) is the Slovenian national library, the central state library and the university library of the University of Ljubljana, the national aggregator of e-content for Europeana and home of the Digital library of Slovenia.

NUK collects, documents, preserves and archives the written cultural and scientific heritage of the Slovenian nation. It provides ready access to knowledge and culture of the past and present Slovenian generations. In collaboration with national and international libraries, it enables access to the world's written cultural and scientific heritage. In the process of creating new knowledge, it helps its users to search, select, evaluate and use information resources in different formats, forms and languages. Its collections and services support scholarly and scientific work of the Ljubljana University and other higher education institutions. The Library is a centre of knowledge aimed at lifelong education of the Slovenian people, and at raising their cultural and educational level and information literacy skills. Through research, development and educational activities in the field of librarianship, the Library is actively co-shaping Slovenian library system, and makes significant contributions to theoretical and practical knowledge of library and information science.

6. Slovenian Language Technologies Society
The Slovenian Language Technologies Society (SDJT) was founded in 1998 and joins people working on language technologies from the scientific, educational or user perspectives. The activities of the SDJT are aimed at promoting the development of language technologies for the Slovenian language. In 2011 the society was awarded special status of a research institution working in the public interest. The society has 120 members.

The main activities of SDJT are its monthly JOTA lecture series on NLP-related topics and the biennial conference on Slovene language technologies. The society also organizes educational events, such as the ESSLLI Summer School on Language, Logic and Information, the TransTech summer school on translation technologies and seminars on corpora and on-line language resources for Slovene for secondary and primary school teachers.

SDJT is a founding member of the CLARIN.SI consortium, and contributes to the goals of CLARIN mainly by outreach activities, such as seminars, tutorials and conference organisation.

7. University of Ljubljana
The University of Ljubljana is the largest Slovenian University, where corpus linguistics and language technologies are coordinated by its Centre for Language Resources and Technologies (CJVT UL). The Centre is a research unit of the University dedicated to scientific research of language, creation and maintenance of practically useful digital language resources and technologies for modern Slovene language, available on the web to all Slovene language users. Research areas include description of modern Slovene language and computer-aided learning and teaching of Slovene and foreign languages. Practical tasks include continuous and user-friendly access to corpora, lexical, terminological and other databases, creation and maintenance of web-based language learning and teaching environments, as well as distribution of publicly financed and open source language resources and tools.

The Centre is organized within the Slovene research agency (ARRS) financed Network of research infrastructure centres at University of Ljubljana. The collaborating partners in the

centre are the Faculty of Social Sciences, Faculty of Arts, Faculty of Education, Faculty of Electrical Engineering, and Faculty of Computer and Information Science.

An important aim of the centre is to provide publicly available information about Slovene language resources and language technologies in Slovenia in order to enhance their public perception, and dissemination of language resources and tools. These aims are strongly related to CLARIN's mission.

8. University of Maribor

The University of Maribor is the second largest and the second oldest university in Slovenia with about 18,000 students. It has seventeen faculties with undergraduate and postgraduate programmes. The University of Maribor is a regional developer and its faculties are located not only in the city of Maribor, but also in other parts of Slovenia. Two of its Faculties are the most active developers and users of language technologies and resources.

The Faculty of Electrical Engineering and Computer Science is devoted to performing education and research within the fields of electrical engineering, computer science, information technology, communications, media, telecommunications and mechatronics. It has contributed to various projects with the development of language resources or language processing tools for Slovene.

The Faculty of Arts is the youngest member of University of Maribor, established in 2006, but with study programs and departments developed in the framework of the Faculty of Pedagogy a long time ago. It covers fields of Slavic studies, English and American studies, German studies, Hungarian studies, history, geography, philosophy, sociology, psychology and pedagogy.

University of Maribor has been a leader or a member of various language technologies projects, thus contributing to the goals of the CLARIN by developing language resources, such as databases for speech recognition or language technologies tools and making them available via the CLARIN.SI repository. Research dealing with language and speech technologies on the University of Maribor is mainly in the domain of the Faculty of Electrical Engineering and Computer Science and its two institutes: Institute of Electronics and Telecommunications and Institute of Computer Science (the laboratory for heterogeneous computer systems). Also, the members of the Faculty of Arts are potential users of the CLARIN services.

9. University of Nova Gorica

The University of Nova Gorica is a young (est. 1995, university accreditation in 2005) and growing private research-oriented university in Slovenia comprising seven schools and twelve research centers. The University is a member of the European University Association (including EUA- Council for Doctoral Education). Despite its still relatively small size (approx. 100 PhD holders), the University of Nova Gorica has hosted numerous nationally- and European-funded research projects including a €4,000,000 FP7-REGPOT grant awarded in 2011, collaborates with over 40 European and international universities and research centers, participates in the academic and research exchange programs (ERASMUS, COST) and in an EC Erasmus Mundus joint study program. It is also member of CLARIN.SI since 2015.

Activities related to CLARIN.SI are conducted in UNG's Center for Cognitive Science of Language. The unit currently employs 6 PhD holders (four senior and two postdoctoral researchers) who specialize in formal theoretical and experimental linguistics, but are also involved in various applied linguistics activities, from conducting language-planning studies for the Slovenian Ministry of Culture to maintaining an online language consultancy. Through

its Center for Cognitive Science of Language, UNG plans to contribute to CLARIN with its language resources.

10. University of Primorska
The University of Primorska covers the Slovenian Littoral region with Faculties and Institutes in Koper, Izola, and Portorož. It has around 5,000 students. Two of its Faculties and one Institute are the most active developers and users of language technologies and resources.

The Faculty for Mathematics, Natural Sciences and Information Technologies offers undergraduate and postgraduate study programmes in mathematics, computer science, natural sciences and biotechnical sciences. Language technologies are mostly used and researched at the Department of Information Sciences and Technologies. Most of the research is done in the fields of Machine Translation and Knowledge discovery.

The Faculty of Humanities offers both undergraduate and postgraduate degree courses as well as engaging in scientific and specialist activities in the field of humanities, arts and social studies. Language technologies are mostly used and researched with the collaboration of the Institute for Linguistic Research of the Science Research Centre. Most of the research is done in the field of corpus-based language studies and corpus construction.

University of Primorska has been a leader or a member of various language technologies projects. The contribution to CLARIN are the provision of domain-specific language corpora and dictionaries and being users of the CLARIN resources and services.

11. Scientific and Research Centre of the Slovenian Academy of Sciences and Arts
The Research Centre of the Slovenian Academy of Sciences and Arts (ZRC SAZU) is the leading Slovenian research centre in the humanities and a cutting-edge academic institution in Central, Eastern, and South-Eastern Europe. It has a multidisciplinary character; in addition to the humanities, its spheres of research also cover the natural and social sciences. It conducts research on a broad variety of topics connected with natural and cultural heritage of Slovenia. In its present form, the ZRC SAZU is a network of eighteen institutes with over 300 researchers and technical specialists.

The largest among the institutes is the Fran Ramovš Institute of the Slovenian Language. It is the national centre for systematic monitoring and description of the Slovenian language. The Institute was established in 1945 for the purpose of compiling linguistic materials and using them for the creation of comprehensive and authoritative Slovenian language resources, primarily dictionaries: orthographic dictionaries, dictionaries of contemporary standard Slovenian, terminological dictionaries, etymological dictionaries, historical dictionaries, dialectal dictionaries, linguistic atlases as well as descriptive and historical studies in linguistics. All monolingual comprehensive dictionaries of Slovenian and many applied ones have been compiled at the Institute of the Slovenian Language. Since 2000, the Institute has published 38 dictionaries on 18,402 pages, 73 monographs on 21,102 pages, and 36 issues of journals on 7785 pages. The majority of these works are also available online with free access.

ZRC SAZU was one of the founding members of CLARIN.SI and contributes to the activities of CLARIN through language resources and consultancy on the Slovenian language.

12. Science and Research Centre Koper
The Science and Research Centre Koper (ZRS Koper) works on an interdisciplinary basis, involving humanities, social and natural sciences with emphasis given to the research in the specific environments of the Mediterranean and the upper Adriatic region. The main activities of the Centre involve basic and applied research, production of professional expertise and

counselling. ZRS Koper is actively integrating into international scientific cooperation and is connecting with many similar organisations worldwide.

Part of ZRS Koper is also the Institute for Linguistic Studies. The Institute focuses on studying the Slovenian linguistic situation in the Northern Adriatic region, which represents a linguistically complex region and a cultural hub between the Central European and Mediterranean regions. The area is discussed from the aspects of theoretical and applied linguistics, as well as literary theory and literary history. Diachronic aspects are considered in studying interferential dialects within the scope of dialectology and etymological research. Synchronous aspects are considered in socio-linguistic research in the field of language politics, Slovene and Italian languages in public use, expert languages and language technologies.

All institutes of ZRS Koper in the field of humanities and social sciences strive to develop permanent and publicly accessible databases as part of their research activities. The activities of the Institute for Linguistic Studies of the ZRS Koper have long focused on the development of text corpora and corpus research. In addition to the freely accessible tourist corpus TURK, we have created several smaller working corpora, such as the corpus of Ivan Cankar's texts and other literary corpora, corpora of media texts on the topic of the pandemic, the corpus of slang texts on Facebook, etc. Aside from adding existing corpora to the CLARIN.SI research infrastructure, we also plan to compile new corpus collections.

The National Coordinator of CLARIN.SI is

Tomaž Erjavec
Dept. of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39
SI-1000 Ljubljana
Slovenia
Email: tomaz.erjavec@ijs.si
Phone: +386 1 477 35 07

The invoice for Annual fee is to be sent to

Ministry of Higher Education, Science and Innovation
ATTN: Dr. Albin Kralj
Masarykova 16
SI-1000 Ljubljana
Slovenia
Email: albin.kralj@gov.si
Phone: +386 1 478 4737

Annex 1-B

The following data and computing centres will be providing services to CLARIN:

- Jožef Stefan Institute (JSI)

B centres and certification status:

The JSI centre was certified as a CLARIN B centre in 2020. It is currently undergoing re-certification.

Annex 1-C

Specification of the implementation of user authentication and authorisation system:

The JSI (and its CLARIN.SI centre) is a part of the AAI federation of the Slovenian Academic and Research Network (ArnesAAI). The repository supports AAI login.

Annex 1-D

National Consortium CLARIN.SI will (i) publish the relevant Knowledge-Sharing Infrastructure activities and facilities it offers on the CLARIN website according to guidelines provided by CLARIN ERIC, (ii) will keep this list up to date, and (iii) will participate in the coordination of these activities with other National Consortia in order to maximize their efficiency and effectiveness.

Annex 1-E

National Consortium CLARIN.SI will contribute to the permanent CLARIN ERIC committees where applicable, and will participate in ad hoc committees as required and within the limits of the national budget:

- National Coordinators' Forum: Tomaž Erjavec (JSI)
- Standing Committee for CLARIN Technical Centres: Cyprian Laskowski (UL)
- Legal and Ethical Issues Committee: Mateja Jemec Tomazin (ZRC SAZU)
- Standards Committee: Tomaž Erjavec (JSI)
- User Involvement Committee: Jakob Lenardič (UL/ERIC)

Any other contribution: none

Annex 1-F

The National Consortium will comply with statutes article 19 on Access Policies for Users.
In certain cases, due to copyright or personal privacy issues some, resources are available only with restricted context or subject to signing of a licence restricting the use to academic context.

Annex 2 National contributions to be coordinated at the national level

Annex 2-A

National Work Programme with timing:

- New on-line annotation tool CLASSLA (2024-02)
- New version of siParl corpus of parliamentary debates (2024-06)
- New version of largest Slovenian corpus metaFida (2024-10)
- Upgrade of storage capacities for CLARIN.SI services (2024-12)
- Update and enhancement of CLARIN.SI web pages (2024-12)
- Installation of new version of Slovene reference corpus Gigafida on CLARIN.SI concordancers (2024-12)
- Continue reviewing and publishing new language resources on repository (2027-12)
- Continue publishing corpora on CLARIN.SI concordancers (2027-12)

Annex 2-B

National Work Programme with timing:

- Deposit of language resource deliverables of the Slovenian Ministry of Culture project "Adaptable processing of natural language with the help of large language models" 2023-2026 (2026-12)
- Deposit of relevant language resources from relevant Slovenian basic research and other projects (2027-12)

Annex 2-C

Measures to promote the adoption of and adherence to CLARIN standards in national resources and tools creation programmes and projects:

- CLARIN.SI promotes the use of CLARIN standards and actively cooperates with submitters to its repository to convert their submission to the appropriate standards
- CLARIN.SI maintains an official XML schema, i.e. a TEI parametrisation (https://github.com/clarinsi/TEI-schema) and encourages users to annotate their resources so that they comply with the schema, e.g. the new generation of reference corpora for Slovene
- Members of CLARIN.SI are leading the work on Parla-CLARIN, a recommendation for encoding parliamentary corpora and producing corpora in accordance with these recommendations.

Annex 2-D

Contributions to the Knowledge Sharing Infrastructure mainly targeting national audiences and not already covered by Annex 1-D:

Occasional lectures on CLARIN given at conferences and other events in Slovenia, c.f. http://www.clarin.si/info/dogodki/

Annex 2-E

Any other contribution: -