CLARIN.SI, the Slovenian node of CLARIN: ten years on

Tomaž Erjavec
Jožef Stefan Institute
ZRC SAZU
tomaz.erjavec@ijs.si

Nikola Ljubešić
Jožef Stefan Institute
University of Ljubljana
nikola.ljubesic@ijs.si

Katja Meden
Jožef Stefan Institute
Inst. of Contemporary History
katja.meden@ijs.si

Taja Kuzman
Jožef Stefan Institute
taja.kuzman@ijs.si

Cyprian Laskowski
University of Ljubljana
cyp@cjvt.si

Jan Jona Javoršek
Jožef Stefan Institute
jona.javorsek@ijs.si

Simon Krek
University of Ljubljana
Jožef Stefan Institute
simon.krek@ijs.si

Mateja Jemec Tomazin
ZRC SAZU
mateja.jemec-tomazin@zrc-sazu.si

Kaja Dobrovoljc
University of Ljubljana
Jožef Stefan Institute
kaja.dobrovoljc@ijs.si

Špela Arhar Holdt
University of Ljubljana
spela.arharholdt@ff.uni-lj.si

Jakob Lenardič
Inst. of Contemporary History
CLARIN ERIC
jakob.lenardic@inz.si

Darja Fišer
Inst. of Contemporary History
CLARIN ERIC
darja.fiser@inz.si

Abstract

The paper presents the organisation and services offered by the Slovenian research infrastructure for language resources and technologies CLARIN.SI. We introduce the governance, organisational structure, and technical components of the infrastructure, followed by a description of its web applications with a focus on the repository and concordancers. Next, we provide an overview of support activities offered by CLARIN.SI, which includes services of the CLASSLA knowledge centre for processing South Slavic languages, financial support for projects, and facilitation of conferences and workshops. We also present the involvement of CLARIN.SI in national and European projects, with its sister national infrastructure nodes of DARIAH and CESSDA, and in the work of CLARIN ERIC.

1 Introduction

The CLARIN.SI consortium was established in 2014, and CLARIN.SI became a member of CLARIN ERIC in 2015. To date, the only publication in English[1] that comprehensively presented CLARIN.SI was published shortly after its establishment (Erjavec et al., 2014), where we described the initial steps of the research infrastructure (RI) and plans for further work. This paper summarises the state of the infrastructure ten years later. In Section 2, we present the organisational structure and management of the RI. In Section 3, we describe the CLARIN.SI repository of language resources and tools. Section 4 presents online services, with a focus on online corpus concordancers, and Section 5 focuses on support activities (knowledge centres, project support, organization of events and other dissemination). In Section 6, we describe the involvement of CLARIN.SI in national and European projects, while Section 7 provides conclusions and plans for further work.

2 Management of CLARIN.SI

The CLARIN.SI infrastructure is based at the Jožef Stefan Institute (JSI), Slovenia's largest research institute. Most of the computer equipment is located at JSI, and the institute also ensures the security, maintenance, and continuous operation of the infrastructure's online services. Three organisational units of the JSI, namely the Department of Knowledge Technologies, the Artificial Intelligence Laboratory and the Centre for Network Infrastructure, are involved in the management and technical maintenance of the infrastructure.

[1] However, see Erjavec et al. (2022a) for a more recent presentation in Slovenian.

2.1 CLARIN.SI Consortium

CLARIN.SI is organised as a consortium, currently comprising 12 partner institutions. The consortium brings together all the key institutions involved in the development or use of language resources and technologies in Slovenia, including research institutes, universities, companies, and an association:

- Research Centre of the Slovenian Academy of Sciences and Arts (ZRC SAZU), in particular its Fran Ramovš Institute for the Slovenian Language, which collects and analyses linguistic materials and produces fundamental works in Slovenian linguistics, especially dictionaries.
- Jožef Stefan Institute (JSI), from where the RI is coordinated, and its repository and most of its services are maintained. The departments involved in the IR also have a strong track record in the development of language resources and tools.
- Institute of Contemporary History (INZ), which coordinates DARIAH-SI and leads the only long-term research programme in Slovenia focused on digital humanities.
- Science and Research Center Koper (located close to the borders with Italy and Croatia), in particular its Institute for Linguistic Studies, which focuses on languages (and cultures) in contact.
- University of Ljubljana, in particular its Centre for Language Resources and Technologies (Center za jezikovne vire in tehnologije, CJVT), which coordinates research in corpus linguistics and language technologies, while developing and maintaining fundamental digital language resources and tools for contemporary Slovenian. The centre is a part of the Network of research and infrastructural centres at the University of Ljubljana, and its activities are carried out at six different faculties, e.g., the Faculty of Computer and Information Science and the Faculty of Arts facilitating interdisciplinary development of written language and speech technologies, and the Faculty of Social Sciences, the Slovenian node of the sister RI for social sciences CESSDA.
- University of Maribor, in particular its Faculty of Electrical Engineering and Computer Science, where language technology and especially speech technology research are performed.
- University of Nova Gorica, in particular its Center for Cognitive Science of Language, which specialises in formal theoretical and experimental linguistics.
- University of Primorska, in particular its Faculty for Mathematics, Natural Sciences and Information Technologies that is active in the fields of machine translation and knowledge discovery.
- National and University Library, which collects, documents, preserves and archives the written cultural and scientific heritage of the Slovenian nation. The Library has joined the consortium only in 2024 and is, of course, an extremely welcome partner in the view of future cooperation (e.g., in transforming parts of its large digital library dLib into language corpora).
- Slovenian Society for Language Technologies (SDJT, est. 1998) promotes the development of language technologies for Slovenian with a focus on open-source solutions. It was for many years the only organiser of the biennial conference series “Language Technologies and Digital Humanities”, and remains the main one.
- Alpineon specialises in developing state-of-the-art computer vision and products involving speech and translation.
- Amebis develops products in the fields of language technology and electronic publishing. It develops text and speech corpora, plug-in language processing modules for Slovenian, and machine translation and speech synthesis systems.

Decisions on the management of the RI are made or confirmed by the CLARIN.SI Management Board, where each partner has one representative and an unlimited number of deputies. Communication occurs through the Board's mailing list, currently comprising 40 subscribers, and during annual meetings where the RI's activities over the past year are reviewed and plans for the following year are formulated.

The operation of the CLARIN research infrastructure in Slovenia is thus based on the needs and consensus of the key stakeholders in digital linguistics and language technologies, as well as digital humanities and social sciences.

2.2 Technical Infrastructure

CLARIN.SI maintains a bilingual (Slovenian, English) website that presents the RI and its services. The website also provides contact information for assistance or advice for users, and password-protected internal pages accessible to members of the Board of Directors, containing founding documents, minutes of meetings, and relevant CLARIN ERIC meeting minutes. In 2024, the CLARIN.SI website received approximately 30,000 visits from users worldwide. As presented in Figure 1, website visits have been consistently increasing over time.

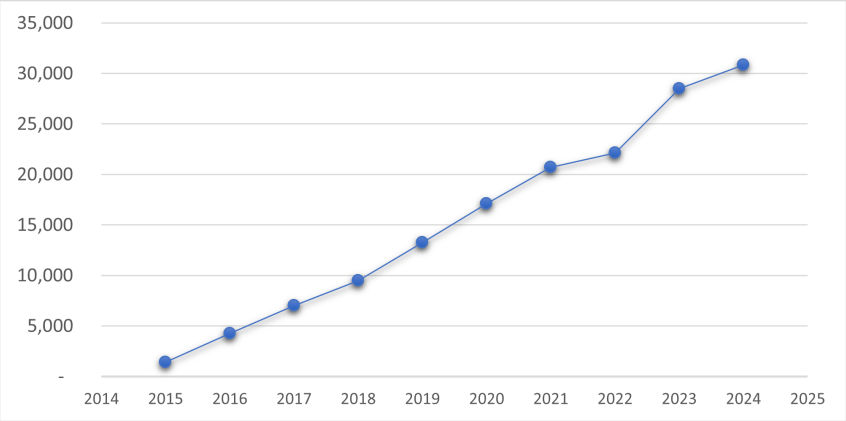

Figure 1: Number of visits to the CLARIN.SI website over time.

Technical documentation is maintained on an internal WordPress installation, where set-up and maintenance procedures for all CLARIN.SI online services are documented. A Redmine installation at the University of Ljubljana is used for managing requests to address identified issues and for planning upgrades.

All changes to critical online services are first tested on the dedicated development servers, where the functioning of the software, documentation and language resources are assessed before deployment on the production servers. The operation of the online services is monitored locally using NAGIOS, while the functioning of the repository is independently verified by the CLARIN ERIC Icinga. In case of errors, service administrators are notified immediately, and can promptly rectify the issue.

3 The CLARIN.SI Repository

The core service offered by CLARIN.SI is its repository of language resources and tools. The repository, set up in 2016, runs on the open-source CLARIN-DSpace platform, which was developed within LINDAT/CLARIAH-CZ at the Institute for Formal and Applied Linguistics at Charles University in Prague. In addition to Slovenia and the Czech Republic, the platform is also used by a number of other national CLARIN repositories, which together represent a third of all regular members of CLARIN ERIC.

3.1 Quality Assurance

Besides ADP, the CLARIN.SI repository is the only one in Slovenia accredited with the Core Trust Seal certificate, which certifies it as a trustworthy data repository. In accordance with the CLARIN ERIC

strategy, the repository implements FAIR principles. The European agenda for open science and the principles of FAIR CLARIN (Jong et al., 2018) are followed using the following instruments:

- AAI academic authentication, which operates using the SSO ("Single sign-on") system. This separates identity providers (e.g., the Slovenian academic network Arnes, the universities and other academic institutions) and service providers (e.g., the repository), allowing users to log into the repository without creating an account on CLARIN.SI. Instead, they use their EduGain username and password with the chosen identity provider.
- Permanent identifiers of entries via the Handle system, which enables the assignment of a permanent URL address to each repository entry.
- Involvement in international online metadata aggregators, such as OpenAIRE, Re3data, and, since 2022, the European Language Grid. CLARIN (and hence CLARIN.SI) was one of the first RIs to be included in the European Open Science Cloud (EOSC). Within CLARIN(.SI), CMDI (Component MetaData Infrastructure) recommendations are followed for metadata records, with export and metadata harvesting supported in the Dublin Core standard.
- A rich selection of licenses, ranging from open ones such as Creative Commons to more restrictive ones that require prior registration and a digital signature of the resource usage agreement.
- Explicit terms of use, which define the rights and obligations of both repository managers and users.
- Instructions for depositing entries, which describe the resource submission process with special emphasis on the required metadata and its format, ensuring uniform and complete metadata records.
- Instructions for encoding deposited data, which specify acceptable record formats and data marking methods, and also include general instructions for the preparation of high-quality and harmonized data. Unlike most other CLARIN repositories (Lenardič & Fišer, 2022b), which typically only list the acceptable formats, CLARIN.SI also offers broader guidance, which is particularly helpful for humanities researchers who may lack advanced computer and data management skills.
- A FAQ about various aspects of the repository and depositing data.

In addition to its main purpose of describing language resources, the CLARIN.SI repository differs from general self-archiving repositories such as Zenodo by ensuring high quality of the deposited language resources and their metadata, as each entry undergoes a careful review by one of the repository editors before publication to ensure it meets the CLARIN.SI criteria. If the criteria are not met, the editor rejects the entry with an explanation of the errors, and, in prearranged cases, assists in correcting the resource.

3.2 Usage and Impact

The repository, at the time of writing, contains 651 entries[2] produced by more than 1,000 authors from over 100 institutions and totalling about 5TB of data. About half of the entries are for or include Slovenian, and about 220 entries include other South Slavic languages (cf. Section 5.1). The top contributors are Jožef Stefan Institute (248 entries), University of Ljubljana (174) and ZRC SAZU (74).

It should be noted that, as a rule, the repository accepts only entries that include data, as we do not want to function merely as a catalogue of language resources, but as their archive. There are two exceptions to this rule. First, if a corpus is mounted on the concordancers offered by the RI but the data cannot be included in the repository (typically because of copyright limitations that prevent download), nor does it have an associated web page showing its authors and other bibliographic information, then we include only the metadata of such a corpus in the repository, in order to enable its users to correctly cite

[2]This number does not include entries hidden for browsing in the repository. We typically hide entries of previous version of submitted resources, as they are not of interest for browsing and merely extend the list of resources. Counting older version of resources as well, the number is closer to 800 entries.

it and access a description of its properties. The second exception is the ELEXIS catalogue of digital dictionaries, as further discussed in Section 6.2.

In Figure 2 we give the number of entries published on the repository since its inception. The number of new entries per year is mostly stable, with about 50 new resources published each year. The exceptions occur in 2022 and 2023, when the ELEXIS collection, the language resources developed in the scope of the national project "Development of Slovene in a Digital Environment" (cf. Section 6.3) and the EU project MaCoCu (cf. Section 6.2), as well as the multilingual CLASSLA annotation models (cf. Section 4.2) were submitted to the repository.

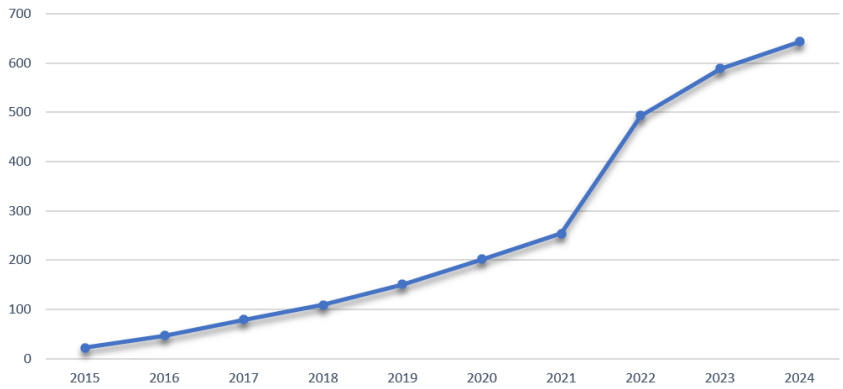

Figure 2: Number of CLARIN.SI repository entries over time.

The number of visits to the CLARIN.SI repository has consistently grown over the years, from about 3,000 in 2015 to 39,000 page views in 2024, although these numbers may include some bots.[3] The most frequently visited resources were the various versions of the ParlaMint corpora, cf. Section 6.5.

As evidenced by the above, CLARIN.SI thus plays a pivotal role in the deposition of open language resources and assists in their creation and description in Slovenia and the region. It has already made a significant contribution to the implementation of the concept of open, verifiable, repeatable and responsible science in the field of linguistic research in Slovenia. It also safeguards numerous language resources created within (Slovenian) research projects from disappearing and provides them with international visibility and influence.

4 Web Services

In addition to the repository, CLARIN.SI maintains several other online services, which are discussed below.

4.1 Concordancers

The most significant services are the concordancers, in particular noSketch Engine and KonText. Both rely on the same back-end program, Manatee (Rychlý, 2007), which enables fast querying of large and richly annotated corpora, the construction of subcorpora, frequency lexica, collocations, etc. To support Slovenian users, CLARIN.SI undertook the localisation of both their user interfaces.

NoSketch Engine is the open-source version of the commercial Sketch Engine concordancer (Kilgarriff et al., 2014), developed by the company Lexical Computing. This is our main concordancer, as it is regularly updated, and has the most advanced functionality and interface. As a matter of principle, we have so far[4] supported anonymous use of the concordancers, so the default installation of noSketch Engine does not require (and therefore support) log-in. This, however, comes with a price, as user-particular

[3] We use Matomo for tracking the use of the repository, and while this platform attempts to exclude bots, it might not always be successful.

[4] Note that this policy might change in the future, as it is becoming increasingly difficult to prevent aggressive harvesting by large companies, which degrades the service.

features are therefore not available (e.g. creation of subcorpora). We therefore recently introduced another instance of noSketch Engine, with self-registration.

The KonText concordancer was developed at the Czech National Corpus Department of Charles University in Prague (Machálek, 2020). It offers a narrower range of functions, but is more in line with the rest of the CLARIN infrastructure: it is used by LINDAT/CLARIAH-CZ, it supports AAI-based log-in, and the currently most recent version should also support FCS.

In addition to these three concordancers, we also still support the old (so-called "Bonito") version of noSketch Engine, though its development was stopped five years ago. However, a number of resources and some services are non-trivially linked directly to this concordancer, such as the glossaries of historical Slovenian linked to the IMP corpus (Erjavec, 2015), the Japanese-Slovenian learner's dictionary of Slovenian (Hmeljak & Erjavec, 2010), linked to several Japanese(-Slovenian) corpora, and the API of the Korpusnik service (cf. Section 4.3).

CLARIN.SI thus maintains four production concordancers, and, in addition, three development instances: old and new noSketch Engine, and KonText. This means that a new corpus has to be mounted on up to seven different virtual machines. To optimise this process, we developed a system of remote management scripts with which, from a remote machine, new (sets of) corpora can be added to any combination of concordancer instances. Additionally, only the development machines are used to index a corpus, which is pertinent for large and heavily annotated corpora, since it is a resource-intensive process that can last over a day. The production machines simply copy all the index files for a corpus to a temporary directory, and then rename it to the production one, minimising disruption of services.

CLARIN.SI concordancers mostly provide access to the same set of corpora, comprising over 200 corpora in 41 languages, including monitor, reference, many types of specialised, and parallel corpora. We here mention only the metaFida corpus (Erjavec, 2023), which combines 34 Slovenian corpora (4.5 billion tokens), making it the largest and most diverse Slovenian corpus available for on-line analysis.

The CLARIN.SI concordances are widely used in university study programmes, linguistic research, research projects, as well as by Slovenian translation companies. They serve a diverse user base, with users originating not only from Slovenia and other South-Slavic-speaking countries, such as Croatia, Serbia, Bulgaria, and North Macedonia, but also from more distant locations, including France, Canada, and Japan.

We have recently implemented tracking of the use of the corpora available on the concordancers. This not only provides us with some key performance indicators, but will also help us to focus on the corpora that are used the most in order to concentrate development where it will have the greatest impact. The analysis is made on the basis of Apache logs, with every effort made to exclude bots.

The analysis of the logs for the new version of noSketch Engine without login, covering the period from June (when tracking was implemented) until December 2024, revealed significant usage of the concordancer, with approximately 400,000 requests in total. Figure 3 presents the daily usage statistics for the recorded period, indicating a consistent demand throughout the year, with an average of 2,000 requests per day.

The analysis showed that users most frequently queried web corpora, particularly the recently developed CLASSLA-web corpora for South Slavic languages (Ljubešić & Kuzman, 2024) and the corpora from the WaCky Web Corpus family (Baroni et al., 2009). In total, web corpora accounted for 65% of all requests made to the top 50 corpora. Other frequently queried corpora included the Slovenian metaFida corpus (Erjavec, 2023); parliamentary corpora, notably those from the ParlaMint collection (cf. Section 6.5); news corpora, such as the Trendi monitor corpus of Slovenian (Kosem, 2022) and the ENGRI corpus of Croatian news portals (Bogunović et al., 2021); the OSS corpus of Slovenian scientific publications (K. Žagar et al., 2023); and the GOS corpus of spoken Slovenian (Zwitter Vitez et al., 2023).

4.2 CLASSLA Annotation Tool

As part of the CLASSLA K-Centre (cf. Section 5), CLARIN.SI also provides the CLASSLA Annotation Tool, an online service for automatic linguistic annotation of raw texts, either by pasting or uploading texts to the web interface.

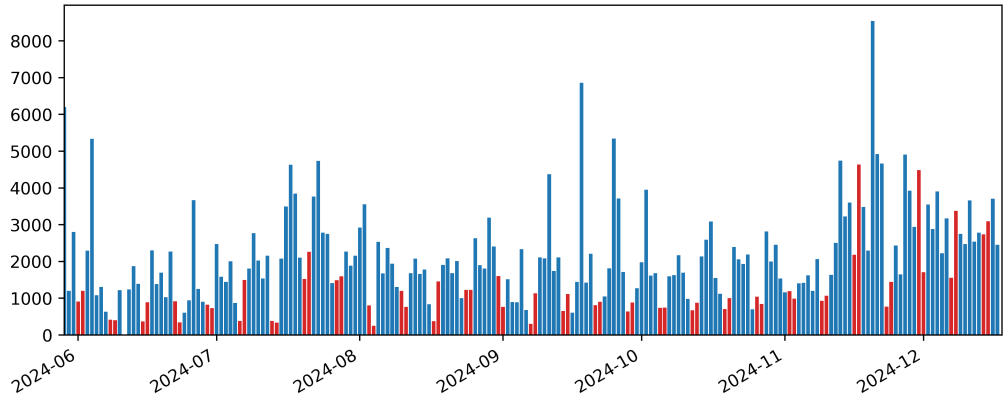

Figure 3: Requests per day through the noSketch Engine in the second half of 2024. Red columns represent requests during weekends.

The web interface, adapted from the CJVT Označevalnik service (Dobrovoljc, 2024b), uses the CLASSLA-Stanza pipeline (Ljubešić et al., 2024d), which is based on the Stanza pipeline (Qi et al., 2020). With the developed models (deposited in the CLARIN.SI repository) the service covers Croatian, Serbian, Slovenian, Macedonian, and Bulgarian, including non-standard (colloquial) varieties of the first three. The tool can annotate the levels of tokenisation, sentence segmentation, morphosyntactic features, lemmas, dependency syntax, named entities, and, for some languages, semantic roles. Additionally, the service enables users to view and download results in various formats, including tables (.xlsx), XML files, CONLL-U files, and graphical images (.svg).

4.3 Korpusnik and SENTA Tools

Partly due to CLARIN.SI-supported projects (cf. Section 5.2) two on-line tools were developed and installed at CJVT, and have been included among the CLARIN.SI services, namely Korpusnik and SENTA tools. Both tools are designed for the general public, with special emphasis on ensuring inclusivity and accessibility.

Korpusnik, a corpus summarizing tool for Slovenian (Kosem et al., 2023) provides a simple and intuitive overview of key information from five Slovenian corpora, selected for their relevance to the general public. This information includes collocations, example sentences, distribution by text type, year of publication, and source. The tool aims to compare and visualise corpus data in a user-friendly way, offering engaging insights into authentic language use.

SENTA (Sentence Simplification and Analysis) is a tool that transforms complex Slovenian sentences into simpler, more comprehensible ones while preserving the original meaning (A. Žagar et al., 2024). The system consists of two components: a neural classifier that identifies sentences requiring simplification and a Slovenian fine-tuned language model based on the T5 architecture to perform the simplification. The tool is accessible through a user-friendly interface, which also provides basic statistics on the texts before and after simplification.

4.4 WebAnno Annotation Platform

The manual corpus annotation platform WebAnno (Yimam et al., 2013), developed within CLARIN-DE, was first used in Slovenia in the context of several national projects, and then an instance was installed at CLARIN.SI, which also manages users, their access levels, as well as the project (annotation) definitions. In the scope of CLARIN.SI we also created scripts to convert TEI-encoded corpora into the WebAnno TSV3 format and to merge manual annotations back into the source TEI corpora (Erjavec et al., 2016).

The more important projects that used CLARIN.SI WebAnno were: "Linguistic analysis of non-standard Slovenian" (Fišer et al., 2020) (manual normalisation of word forms and for assigning lemmas and morphological annotation to Slovenian user-generated content); "Slovenian scientific texts: sources and description" (Erjavec et al., 2021) (marking bilingual terms); "Terminology and knowledge schemes in the interlinguistic space" (Vintar & Martinc, 2022) (marking term definitions in texts); "May '68 in Literature and Theory: The Last Season of Modernism in France, Slovenia, and the World" (marking named entities, and foreign and non-standard language words and phrases) (Žejn & Šorli, 2023); the EU project "INTAVIA: In/Tangible European Heritage, Visual Analysis, Curation & Communication"(marking up abbreviations in the Slovenian Biographical Lexicon) (Daza et al., 2022); and a series of projects aimed at manual syntactic parsing of written and spoken Slovene (Dobrovoljc, 2024a; Dobrovoljc et al., 2023).

4.5 Usage of Git

CLARIN.SI also leverages Git platforms for collaborative development of software and language resources. On GitHub, CLARIN.SI hosts over 100 open-source projects under its virtual organisation. For projects requiring local hosting or restricted access due to copyright concerns, CLARIN.SI also operates a GitLab installation. This platform hosts around 20 projects, both public (such as the already mentioned TEI conversion for WebAnno) and private (mostly related to CLARIN.SI services).

5 Support and Dissemination

Apart from the repository and web services offered by CLARIN.SI, the infrastructure also supports the user community via the knowledge centres it belongs to, and by an annual call for projects, as well as by disseminating its results, as discussed in this section.

5.1 Knowledge Centres

CLARIN.SI is active in promoting and encouraging the development of computational linguistics not only for Slovenian but also for other South Slavic languages (i.e., Croatian, Serbian, Bosnian, Montenegrin, Macedonian and Bulgarian), significantly increasing the international use of the RI.

Together with the Bulgarian CLARIN research infrastructure (CLADA-BG) and the Institute of Croatian Language, CLARIN.SI manages the CLARIN Knowledge Centre for South Slavic languages CLASSLA, which provides assistance in the use of language resources and technologies for the whole South Slavic language group. CLASSLA supports researchers with documentation on open language resources, tools for creating and processing text corpora, and other language technologies. In 2024, the centre also organised a series of tutorials called CLASSLA-Express, which is described in detail in Section 5.4.

In addition, the CLASSLA centre develops its own language technologies and corpora to meet the needs of South Slavic languages. Notable resources, developed by the knowledge centre, are the CLASSLA-Stanza models for linguistic annotation (Ljubešić et al., 2024d), discussed in Section 4.2, many linguistically annotated text corpora, including the first general linguistically annotated corpus of Macedonian, CLASSLA-web.mk (Ljubešić et al., 2024c); and the first large and open speech corpora for Croatian and Serbian called ParlaSpeech (Ljubešić et al., 2024b). In 2024, the CLASSLA knowledge centre set up a crawling infrastructure for the annual collection of web corpora for South Slavic languages. The first version of corpora, CLASSLA-web 1.0 (Ljubešić & Kuzman, 2024), comprises 11 billion words in 7 languages, which represent the largest general corpora for most South Slavic languages. The corpora, which are automatically annotated with linguistic and genre information, are available both on the CLARIN.SI concordancers and the CLARIN.SI repository, enabling linguistic research as well as development of language technologies on these languages.

In 2021, CLARIN.SI also became a member of the CLARIN Knowledge Centre for Processing User-Mediated Communication CKCMC, managed by Eurac Research in Bolzano, while in 2024 it established the CLARIN-ELEXIS Knowledge Centre for Lexicography, which operates as a distributed virtual centre supported by 17 institutions from 12 CLARIN National Consortia. It offers expertise and support in using

open-access data, tools and services for lexicographers and is a follow-up of the ELEXIS (European Lexicographic Infrastructure) H2020 project.

5.2 Project Support

CLARIN.SI provides financial support for projects selected through an annual call open to consortium members. Since the initiative began in 2018, 36 projects have been successfully concluded, producing notable results, such as the first version of the corpus of parliamentary debates of the National Assembly of the Republic of Slovenia, siParl, now at version 4 (Pančur et al., 2024), the speech corpus Gos Video-Lectures (Verdonik et al., 2019), the LIST tool for analysis of Slovenian corpora (Krsnik et al., 2019), the SloBENCH evaluation framework for language technologies (Žitnik & Dragar, 2021), and teaching materials for corpus approaches to the research of parliamentary discourse (Fišer & Pahor de Maiti, 2021).

5.3 Organization of Events

CLARIN.SI supports events in the field of computational linguistics and related topics that take place in Slovenia, such as the recent 34th European Summer School in Logic, Language and Information (ESSLLI 2023, University of Ljubljana, Faculty of Computer and Information Science) and SyntaxFest 2025 (Faculty of Law, University of Ljubljana).

CLARIN.SI is one of the organisers of the International Conference on Language Technologies and Digital Humanities, a biennial event held in Ljubljana and the primary conference for the field in Slovenia. The conference, first held in 1998 (then called "Language Technologies"), was initiated by the SDJT association, as mentioned in Section 2.1, one of the members of the CLARIN.SI consortium. The conference features invited lectures, on-line proceedings, a student session and associated events, such as tutorials, panels and workshops.

Since 2005, SDJT has been organising a series of lectures, known as JOTA, where CLARIN.SI supports the recording and archiving of lectures on VideoLectures.NET. So far, 20 lectures have been recorded and viewed 15,000 times.

5.4 Promotion and Training

CLARIN.SI presents its activities and those of its knowledge centres at national and international workshops, conferences and events, such as the ESFRI and CLARIN conferences. The work of CLARIN.SI and the CLASSLA K-centre was presented as part of the "Tour de CLARIN" initiative in 2019 (Fišer et al., 2019).

The RI organises training sessions on compiling, depositing and using corpora and other language resources (e.g., using the noSketch Engine concordancer, WebAnno, and Git platforms). In particular, CLASSLA participated in a workshop on using corpora for analysis of the regional variation of gender marking in a language, and, in 2024, organised the CLASSLA-Express series of tutorials, which comprised hands-on exercises on using the CLASSLA-web corpora on the CLARIN.SI concordancers (Ljubešić et al., 2024a). Six CLASSLA-Express training sessions took place so far: Croatia (Zagreb and Rijeka), Serbia (Belgrade), North Macedonia (Skopje), Bulgaria (Sofia), and Slovenia (Ljubljana). The second iteration of CLASSLA-Express in 2025, focusing on comparing traditional corpus-based approaches to approaches using large language models, is visiting five cities in three countries, showing that the workshop series has potential to become a continuous event.

CLARIN.SI shares regular updates on the activities of the Consortium members and Knowledge Centres through its website and the CLASSLA mailing list, with about 100 subscribers from Slovenia and abroad, including Croatia, Serbia, Montenegro, Bulgaria, North Macedonia, Italy, Spain, France, Germany, and the USA.

The infrastructure has accounts and posts news on the social media platforms Discord (100 members), LinkedIn (150 followers) and X (270 followers).

6 Involvement in Projects

CLARIN.SI participates in national and European projects, thus promoting the utilisation of its resources, and enhancing its visibility.

6.1 European Cohesion Policy Funds

The Slovenian 2018–2021 cohesion projects financed upgrading the research equipment of ESFRI infrastructures. With these funds three CLARIN.SI consortium partners upgraded their computer equipment. JSI significantly upgraded its computer cluster, with high-end machines and plentiful disk storage, and likewise CJVT, while the University of Maribor acquired an Nvidia GPU server, which is used for research into deep learning of big data language processing. With these upgrades, CLARIN.SI provides the Slovenian research community with a state-of-the-art research infrastructure, attracting Slovenian partners to international research and innovation projects and supporting scientific excellence. For instance, the EU project MaCoCu (Bañón et al., 2022) used the CLARIN.SI cluster to collect and process large web corpora for over 10 European languages.

6.2 Involvement in European Projects

While CLARIN.SI has not been directly involved in EU projects, there were several language-related projects that were either led by Slovenian researchers or Slovenian teams participated in them, and these projects deposited the language resources they developed in the CLARIN.SI repository (e.g., the MaCoCu web corpora (Bañón et al., 2022), and the news corpora for several languages of the EMBEDDIA project).

As one of its goals, the EU https://elex.is/ELEXIS project had set the creation of a catalogue of online European dictionaries, which was deposited as a dedicated collection into the CLARIN.SI repository. The collection contains metadata and links to the on-line portals of 143 digital dictionaries.

6.3 Involvement in National Projects

CLARIN.SI directly or indirectly participated in numerous national projects, the largest being "Development of Slovenian in a Digital Environment". CLARIN.SI contributed by reviewing language resources created within the project, which were then deposited in its repository, and by defining schemas for the mark-up of Slovenian language resources.

6.4 Cooperation with other infrastructures and initiatives

CLARIN.SI works closely with its two sister RIs in Slovenia, namely DARIAH-SI, the national RI node for digital humanities, and ADP, the national RI node of the CESSDA, the Consortium of European Social Science Data Archives. As already mentioned in Section 2.1, both RIs are led by institutions that are members of the CLARIN.SI consortium.

CLARIN.SI has a long-standing collaboration with DARIAH-SI in the development of encoding and creation of parliamentary corpora, starting with the creation of Slovenian parliamentary corpora (Meden et al., 2024), and continued in the Parla-CLARIN and ParlaMint initiatives, as further explained in Section 6.5. With ADP we collaborated in the project "RDA Node Slovenia" (2019–2020), in the scope of which we established the national RDA Node, which acts as a long-term contact point between the Research Data Alliance and Slovenian data practitioners. In this context CLARIN.SI also reviewed and analysed Slovenian research data repositories (Meden & Erjavec, 2021).

CLARIN.SI is also a founding member of the Slovenian national supercomputer network SLING and of the recently established Slovenian Open Science Community.

6.5 Participation in the Work of CLARIN ERIC

CLARIN.SI plays an active role in the work of CLARIN ERIC, not only by participating in its various bodies, but also by its members directly contributing to the ERIC and by leading CLARIN projects, as evident by the various awards received by its members. In addition, representatives of the CLARIN.SI

management committee participate in the CLARIN committees on Legal issues, Standardisation, User involvement, and Technical Centres.

Jakob Lenardič (INZ) received the CLARIN Steven Krauer Award for the young researcher of the year in 2019, also for his work (together with Darja Fišer) in establishing the CLARIN Resource Families initiative (Lenardič & Fišer, 2022a). Tomaž Erjavec received the Steven Krauer Award for CLARIN Achievements 2021 for his work on the ParlaMint project, Darja Fišer (INZ) and Kristina Pahor de Maiti Tekavčič (UL, INZ) received the Teaching with CLARIN Award in 2021 for the best teaching material related to the use of CLARIN resources, while Ajda Pretnar Žagar (UL, INZ), Kristina Pahor de Maiti Tekavčič (UL, INZ) and Darja Fišer (INZ) received the 2022 Teaching with CLARIN Award for their tutorial "What's on the Agenda? Topic Modelling Parliamentary Debates before and during the COVID-19 Pandemic". Kaja Dobrovoljc (UL, JSI) presented CLARIN.SI at the conference on the 20th anniversary of ESFRI in Paris in 2022.

Between 2016 and 2020, Darja Fišer was the CLARIN director for user involvement, and since 2023 she has been the executive director of CLARIN ERIC.

CLARIN.SI (JSI) led two CLARIN projects that included international workshops in 2016 (Ljubljana) and 2019 (Amersfoort). The latter, in cooperation with DARIAH-SI, was dedicated to the development of recommendations for the standardised coding of corpora of parliamentary debates under the name Parla-CLARIN (Erjavec & Pančur, 2022), which has become a popular choice for encoding parliamentary corpora. On this basis, CLARIN.SI acquired a key role in two major CLARIN Flagship projects, ParlaMint I (2020–2021) and ParlaMint II (2022–2023).

The ParlaMint projects created comparable, interpretable and uniformly coded corpora of parliamentary debates and an environment to compile, convert, and process the corpora. In ParlaMint I, CLARIN.SI led the collection and encoding of 17 corpora of national parliaments (Erjavec et al., 2022b). ParlaMint II expanded and enriched the existing corpora while also adding new ones, and resulted in the production of 29 corpora (Erjavec et al., 2024). CLARIN.SI members (co-)led four of the five work packages of the project. In both projects, the corpora were deposited to the CLARIN.SI repository and mounted on its concordancers. Each project released the corpora in three versions (pilot, project final, bug fix) and the corpora are available in several variants (text, linguistically analysed, and machine translated to English in the more recent versions). Additionally, ParlaSpeech corpora consisting of subsets of ParlaMint corpora with aligned speech (Ljubešić et al., 2024b) were released for four languages, Croatian (already in version 2), Serbian, Polish and Czech, spanning five thousand hours of recorded material. Currently, an extension of the ParlaSpeech corpora collection is being prepared, which includes the Slovenian, Bulgarian, Bosnian, Serbian, and Ukrainian ParlaMint dataset. The repository thus currently hosts 19 ParlaMint entries, while the concordancers (which have each country as a separate corpus + joint parallel original-MTed corpus + three ParlaSpeech corpora) mount 139 ParlaMint corpora.

The results of the ParlaMint project are now being used in two successor projects. The on-going OSCARS project ParlaCAP "Comparing agenda settings across parliaments via the ParlaMint dataset´´, which we lead, aims to enhance the speeches in ParlaMint corpora with information about sentiment expressed and topic discussed, add information about political parties from new resources, and especially open this dataset for political scientists and other researchers that rely on tabular data rather than on textual collections. The project PressMint "Interoperable corpora of historical newspapers" has been accepted for funding in the 2025 CLARIN Flagship project call, and will take the developed ParlaMint infrastructure and apply it to compiling a set of historical newspaper corpora. In this project we lead a work-package and several tasks and will contribute the Slovenian corpus.

7 Conclusions

The paper has presented the Slovenian CLARIN infrastructure in its tenth year of existence. The focus has been on the management of CLARIN.SI, its repository for language resources and tools, the web services it offers, its contributions to dissemination activities and support of the field in Slovenia, and its involvement in various projects and in the work of CLARIN ERIC. The overview shows that CLARIN.SI is an established infrastructure that covers a wide interdisciplinary field and supports both basic and

applied research, as well as the development of resources and tools.

As regards further work, the general directions are given in the CLARIN.SI strategy 2024–2030, which was written and approved by the CLARIN.SI Management Committee at the end of 2023. The strategy follows the CLARIN ERIC Strategy 2024–2026 but is focused on the Slovenian node of CLARIN and on cooperation of CLARIN.SI with and within CLARIN ERIC. The CLARIN.SI Strategy is aligned with Slovenia's Research Infrastructure Roadmap 2030, and thus also covers a longer time period than the CLARIN Strategy.

Specifically, we plan in the next period, first and foremost, to maintain and support existing services, thus justifying our status as an infrastructure. Here the main challenge we face is upgrading our current repository platform to the new version being developed by LINDAT/CLARIAH-CZ, which will involve a complex migration procedure of the documentation and URLs.

Next, CLARIN.SI will continue to promote the production of FAIR research data as well as data reuse, especially among humanities researchers. This will necessitate strengthening user support, including education and training, especially as universities and research agencies are increasingly demanding from researchers in doctoral and research programs plans for handling research data and their permanent storage. We plan to adopt a more proactive approach to conducting lectures and workshops for students, lecturers and researchers (as piloted e.g., by the CLASSLA-Express series of tutorials), to provide assistance in creating a data management plan for students and researchers, and to extend our online documentation and tutorials. A survey is currently being prepared to gather feedback on user experience with the CLARIN.SI infrastructure. The survey aims to identify key issues and priorities as seen by the users, helping to harmonise future work with user needs.

There is also the growing importance of research infrastructures for capturing, storing and processing data from social networks and the web. CLARIN.SI has already paid special attention to such language resources, and in the future, it will continue these activities for all South Slavic languages within the CLASSLA knowledge centre.

It is, furthermore, important to instrumentalise and make accessible data and services relevant to individual communities. The CLARIN.SI consortium currently includes 12 members, which cover the majority of Slovenian stakeholders who either produce or use language resources and technologies, but not all of them. In the next period, CLARIN.SI will try to expand its consortium to also cover communities of potential users of the infrastructure that have not yet been included in its operation.

Last but not least, Slovenia's Research Infrastructure Roadmap 2030 states that Slovenia "plans to continue and strengthen activities within the framework of international CLARIN projects" (p. 60), acknowledges the existing cooperation with RI DARIAH-SI and CESSDA/ADP, and foresees connection with new RIs, namely OPERAS (Open scientific communication in the European research area for social sciences and humanities), which is managed in Slovenia by ZRC SAZU, and with PRACE (Partnership for Advanced Computing in Europe), managed by ARNES, the Academic and Research Network of Slovenia.

Acknowledgements

We would like to thank the three anonymous reviewers for their helpful suggestions. The presented work was supported by the Slovenian research agency ARIS in the scope of its funding for ESFRI research infrastructures.

References

Bañón, M., Esplà-Gomis, M., Forcada, M. L., García-Romero, C., Kuzman, T., Ljubešić, N., van Noord, R., Sempere, L. P., Ramírez-Sánchez, G., Rupnik, P., et al. (2022). MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. *23rd Annual Conference of the European Association for Machine Translation*, 301–302. https://aclanthology.org/2022.eamt-1.41/

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, *43*, 209–226. https://doi.org/10.1007/s10579-009-9081-4

Bogunović, I., Kučić, M., Ljubešić, N., & Erjavec, T. (2021). *Corpus of Croatian news portals ENGRI (2014-2018)*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1416

Daza, A., Fokkens, A., & Erjavec, T. (2022). Dealing with abbreviations in the Slovenian biographical lexicon. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 8715–8720. https://aclanthology.org/2022.emnlp-main.596/

Dobrovoljc, K. (2024a). Extending the Spoken Slovenian Treebank. *Proceedings of the Conference on Language Technologies and Digital Humanities*, 113–143. https://doi.org/10.5281/ZENODO.13936394

Dobrovoljc, K. (2024b). Can't See the Forest for the Trees: Tools and Services for Investigating Slovene Dependency Treebanks. *Proceedings of the CLARIN Annual Conference*. https://www.clarin.eu/sites/default/files/CLARIN2024_ConferenceProceedings_final.pdf

Dobrovoljc, K., Terčon, L., & Ljubešić, N. (2023). Universal dependencies za slovenščino. *Slovenščina 2.0*, *11*(1), 218–246. https://doi.org/10.4312/slo2.0.2023.1.218-246

Erjavec, T. (2015). The IMP historical Slovene language resources. *Language Resources and Evaluation*, *49*, 753–775. https://doi.org/10.1007/s10579-015-9294-7

Erjavec, T. (2023). *Corpus of combined Slovenian corpora metaFida 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1775

Erjavec, T., Dobrovoljc, K., Fišer, D., Javoršek, J. J., Krek, S., Kuzman, T., Laskowski, C. A., Ljubešić, N., & Meden, K. (2022a). Raziskovalna infrastruktura CLARIN.SI (The CLARIN.SI research infrastructure). *Proceedings of the Conference on Language Technologies and Digital Humanities*, 47–54. https://doi.org/10.5281/zenodo.14165471

Erjavec, T., Fišer, D., & Ljubešić, N. (2021). The KAS corpus of Slovenian academic writing. *Language Resources and Evaluation*, *55*(2), 551–583. https://rdcu.be/b7GrB

Erjavec, T., Holdt, Š. A., Čibej, J., Dobrovoljc, K., Fišer, D., Laskowski, C., & Zupan, K. (2016). Annotating CLARIN.SI TEI corpora with WebAnno. *Proceedings of the CLARIN annual conference*. https://www.clarin.eu/sites/default/files/erjavec-etal-CLARIN2016_paper_17.pdf

Erjavec, T., Javoršek, J. J., & Krek, S. (2014). Raziskovalna infrastruktura CLARIN.SI. *Zbornik Devete konference JEZIKOVNE TEHNOLOGIJE*. https://nl.ijs.si/isjt14/proceedings/isjt2014_03.pdf

Erjavec, T., Kopp, M., Ljubešić, N., Kuzman, T., Rayson, P., Osenova, P., Ogrodniczuk, M., Çöltekin, Ç., Koržinek, D., Meden, K., Skubic, J., Rupnik, P., Agnoloni, T., Aires, J., Barkarson, Starkaður, Bartolini, R., Bel, N., Pérez, C. M., … Fišer, D. (2024). ParlaMint II: Advancing Comparable Parliamentary Corpora Across Europe. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-024-09798-w

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M. C., de Macedo, L. D., Navarretta, C., Luxardo, G., … Fišer, D. (2022b). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-021-09574-0

Erjavec, T., & Pančur, A. (2022). The Parla-CLARIN recommendations for encoding corpora of parliamentary proceedings. *Journal of the Text Encoding Initiative (Selected Papers from the 2019 TEI Conference)*, (14). https://doi.org/10.4000/jtei.4133

Fišer, D., Lenardič, J., Auziņa, I., Bernstein Ratner, N., De Smedt, K., Dobrovoljc, K., Dodé, R., Domeij, R., Dyvik, H., Erjavec, T., Gerassimenko, O., Hajič, J., Křen, M., Ljubešić, N., MacWhinney, B., Monachini, M., Nava, B., Navarreta, C., Nedyalkova, A., … Vider, K. (2019). *Tour de CLARIN Volume Two*. Zenodo. https://doi.org/10.5281/zenodo.3754164

Fišer, D., Ljubešić, N., & Erjavec, T. (2020). The Janes project: Language resources and tools for Slovene user generated content. *Language Resources and Evaluation*, *54*, 223–246. https://rdcu.be/7RX4

Fišer, D., & Pahor de Maiti, K. (2021). "Prvič, sem političarka in ne politik, drugič pa…": Korpusni pristop k raziskovanju parlamentarnega diskurza. *Contributions to Contemporary History*, *61*(1). https://doi.org/10.51663/pnz.61.1.07

Hmeljak, K., & Erjavec, T. (2010). The Japanese-Slovene dictionary jaSlo: its developments, enhancement and use. *Studia Kognitiva*, *10*, 211–224. https://nl.ijs.si/jaslo/bib/HmeljakErjavec2010.pdf

Jong, F. D., Maegaard, B., Smedt, K. D., Fišer, D., & Uytvanck, D. V. (2018). CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. https://aclanthology.org/L18-1515

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*, 7–36. https://www.sketchengine.eu/wp-content/uploads/The_Sketch_Engine_2014.pdf

Kosem, I. (2022). Trendi - a monitor corpus of Slovene. *Dictionaries and Society. Proceedings of the XX EURALEX International Congress, 12-16 July 2022, Mannheim, Germany*, 230–239. https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202022/EURALEX2022_Pr_p1-21_Front-matter.pdf

Kosem, I., Čibej, J., Krek, S., & Dobrovoljc, K. (2023). Korpusnik: a corpus summarizing tool for Slovene. *CLARIN Annual Conference Proceedings*, 129. https://office.clarin.eu/v/CE-2023-2328_CLARIN2023_ConferenceProceedings.pdf

Krsnik, L., Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Ključevšek, A., Krek, S., & Robnik-Šikonja, M. (2019). *Corpus extraction tool LIST 1.2*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1276

Lenardič, J., & Fišer, D. (2022a). The CLARIN resource and tool families. In D. Fišer & A. Witt (Eds.), *CLARIN. The infrastructure for language resources* (pp. 343–372). https://doi.org/10.1515/9783110767377-013

Lenardič, J., & Fišer, D. (2022b). CLARIN Depositing Guidelines: State of Affairs and Proposals for Improvement. *Proceedings of the CLARIN Annual Conference*. https://www.clarin.eu/event/2022/clarin-annual-conference-2022

Ljubešić, N., & Kuzman, T. (2024). CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 3271–3282. https://aclanthology.org/2024.lrec-main.291/

Ljubešić, N., Kuzman, T., Petrović' Filipović', I., Parizoska, J., & Osenova, P. (2024a). CLASSLA-Express: a Train of CLARIN.SI Workshops on Language Resources and Tools with Easily Expanding Route. *CLARIN Annual Conference Proceedings*, 31. https://doi.org/10.48550/arXiv.2412.01386

Ljubešić, N., Rupnik, P., & Koržinek, D. (2024b). The ParlaSpeech collection of automatically generated speech and text datasets from parliamentary proceedings. *International Conference on Speech and Computer*, 137–150. https://doi.org/10.1007/978-3-031-77961-9_10

Ljubešić, N., Rupnik, P., & Kuzman, T. (2024c). *Macedonian web corpus CLASSLA-web.mk 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1932

Ljubešić, N., Terčon, L., & Dobrovoljc, K. (2024d). CLASSLA-Stanza: the next step for linguistic processing of South Slavic languages. *Proceedings of the Conference on Language Technologies and Digital Humanities*, 251–274. https://zenodo.org/records/13936406

Machálek, T. (2020). KonText: Advanced and Flexible Corpus Query Interface. *Proceedings of the 12th Language Resources and Evaluation Conference*, 7003–7008. https://www.aclweb.org/anthology/2020.lrec-1.865

Meden, K., & Erjavec, T. (2021). *Pregled slovenskih repozitorijev raziskovalnih podatkov* (tech. rep.). Jožef Stefan Institute. CLARIN.SI. https://www.clarin.si/info/services/projects/%5C#RDA_Node_Slovenia

Meden, K., Erjavec, T., & Pančur, A. (2024). Slovenian parliamentary corpus siParl. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-024-09746-8

Pančur, A., Meden, K., Erjavec, T., Ojsteršek, M., Šorn, M., & Blaj Hribar, N. (2024). *Slovenian parliamentary corpus (1990-2022) siParl 4.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1936

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://doi.org/10.48550/arXiv.2003.07082

Rychlý, P. (2007). Manatee/Bonito - A Modular Corpus Manager. *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, 65–70. https://nlp.fi.muni.cz/raslan/2007/papers/12.pdf

Verdonik, D., Potočnik, T., Sepesy Maučec, M., Erjavec, T., Majhenič, S., & Žgank, A. (2019). *Spoken corpus Gos VideoLectures 4.0 (transcription)*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1223

Vintar, Š., & Martinc, M. (2022). Framing karstology: From definitions to knowledge structures and automatic frame population. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *28*(1), 129–156. https://doi.org/10.1075/term.21005.vin

Yimam, S. M., Gurevych, I., Eckart de Castilho, R., & Biemann, C. (2013). WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 1–6. https://aclanthology.org/P13-4001

Žagar, A., Klemen, M., Robnik-Šikonja, M., & Kosem, I. (2024). SENTA: Sentence simplification system for Slovene. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 14687–14692. https://aclanthology.org/2024.lrec-main.1279/

Žagar, K., Ferme, M., Ojsteršek, M., Jemec Tomazin, M., & Erjavec, T. (2023). *Corpus of scientific texts from the Open Science Slovenia portal OSS 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1774

Žejn, A., & Šorli, M. (2023). Named Entities in Modernist Literary Texts: The Annotation and Analysis of the May68 Corpus. *Slovenščina 2.0*, *11*(1), 118–137. https://doi.org/10.4312/slo2.0.2023.1.118-137

Žitnik, S., & Dragar, F. (2021). *SloBENCH evaluation framework*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1469

Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., Erjavec, T., Verdonik, D., Potočnik, T., Sepesy Maučec, M., Majhenič, S., Žgank, A., Bizjak, A., Gril, L., Dobrišek, S., Križaj, J., Bajec, M., Lebar Bajec, I., Jelovšek, T., Trojar, M., Bernjak, M., … Dobrovoljc, K. (2023). *Spoken corpus Gos 2.0 (transcriptions)*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1771