

Users' Experience and Development Priorities for CLARIN.SI

Špela Arhar Holdt

Faculty of Computer and Information Science

& Faculty of Arts,

University of Ljubljana, Slovenia

arharhs@ff.uni-lj.si

Katja Meden and **Taja Kuzman Pungeršek**

Jožef Stefan Institute,

Jožef Stefan International Postgraduate School,

Ljubljana, Slovenia

{katja.meden,taja.kuzman}@ijs.si

Nikola Ljubešić

Jožef Stefan Institute;

Faculty of Computer and Information Science,

University of Ljubljana; Institute of Contemporary History,

Ljubljana, Slovenia

nikola.ljubesic@ijs.si

Tomaž Erjavec

Jožef Stefan Institute, Ljubljana, Slovenia

tomaz.erjavec@ijs.si

Abstract

This paper reports the results of a user survey evaluating the services, tools, and broader functioning of the CLARIN.SI research infrastructure. The survey was designed to identify areas for improvement and to support development planning and strategic decision-making for the coming years. The questionnaire included 50 questions across eight thematic sections, covering the full range of user engagement with the infrastructure, from depositing and downloading resources to using tools and attending events. It was distributed to existing and potential users in Slovenia and abroad. Based on 228 responses, the results show that CLARIN.SI is generally highly valued by its community, especially for the reliability of its data, its support for open science, and the usefulness of its resources. The findings also point to a highly interdisciplinary user base and to a substantial group of potential users who have not yet directly engaged with the infrastructure. Although repository services and language tools are widely used and positively evaluated, respondents consistently identified challenges related to resource discovery, metadata clarity, and search functionality. More broadly, the results suggest that the future impact of CLARIN.SI will depend not only on technical development, but also on effective outreach, onboarding, and training, including the provision of concrete use cases and reusable educational materials.

1 Introduction

CLARIN.SI, the Slovenian node of the CLARIN infrastructure, recently celebrated its 10th anniversary. Since its establishment, the infrastructure has engaged in a range of activities, both national and international. An overview of its organisation, main areas of activity, and impact on the research community is presented in Erjavec et al. (2025).

To mark this milestone and support planning for the next decade, we set out to gather user feedback that would highlight what works well, what needs improvement, and how we might reach new users who are not yet familiar with the infrastructure. With these goals in mind, we designed a survey, focusing on the quality of the tools, resources, and services offered by CLARIN.SI. To engage both current and potential users, we shared the survey not only within our usual networks but also more broadly, aiming to reach as wide an audience as possible, both in Slovenia and abroad.

Beyond its immediate role in informing the future development of CLARIN.SI, the survey can also be situated within broader practices of research infrastructure evaluation. In distributed infrastructures such as CLARIN, user feedback is important not only for improving services, but also for setting priorities,

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

planning sustainable development, and aligning infrastructure provision with the needs of its research communities. From this perspective, the present study offers a structured account of user views on the accessibility, usefulness, and future development of CLARIN.SI services. It may therefore be of interest not only in the Slovenian context, but also to other CLARIN centres and related digital research infrastructures seeking to support planning and evaluation through direct user consultation.

It is noteworthy that comparable studies, at the moment, seem relatively scarce. Italy reported on a survey (Nicolas et al., 2018) carried out only two years after the establishment of CLARIN-IT, which was restricted to a sample of Italian digital humanists interested in ancient Greek philology. While the survey is not presented in detail, the main conclusions were that most of the available resources did not correspond to users' requirements which are centred around authoring, editing, indexing and presenting a digital edition, and the ability to link the available resources and improve them, i.e. they would need tools to integrating textual data and bibliography links. This survey thus targeted a very specific user community, which is well outside the focus of CLARIN.SI; in Slovenia, publishing digital editions is, by agreement, left to the DARIAH-SI infrastructure, while CLARIN.SI concentrates on corpus download and analysis. The other survey was carried out in Norway (De Smedt, 2020), and, while the questions were somewhat more limited than our survey, the ones given were similar to ours, except for services offered by CLARINO rather than CLARIN.SI. As will be seen in the conclusions, their summary is quite similar to ours: CLARINO had, at the time, satisfied to very satisfied users, although some had unrealistic expectations from the infrastructure. It also appeared that some did not realize what CLARINO is, with the lesson that better information to the target user group is at least as important as maintaining and improving the infrastructure services.

A related landscape survey was conducted by the Humanities and Cultural Heritage Open Science Cloud (H2IOSC), featuring the collaborative involvement of Italian ESFRI research infrastructures (CLARIN, DARIAH, E-RHIS, and OPERAS). The survey questionnaires represent only part of their overall monitoring activities, which also include interviews, focus groups, and landscape mapping, among others (Sichera et al., 2025). While the questionnaire design consists of at least somewhat comparable sections, the results pertain to the overall state of the H2IOSC federation, targeting a more diverse and heterogeneous community and providing an overview of community behaviours and needs, rather than direct feedback. The results (Luzietti et al., 2025; Monachini et al., 2025) indicate stable awareness and involvement among users from diverse disciplinary fields, with a wide selection of tools and services used in their work and strong interest in access to archives and repositories. The results are therefore less comparable with our own survey; however, the work and activities accompanying the survey could be adapted for CLARIN.SI and other Slovenian research infrastructures for in-depth landscape analysis.

2 Method

User opinions were collected through an online survey written in Slovenian and English language, designed and administered using the 1KA survey platform. The questionnaire consisted of 50 questions divided into 8 thematic sections. To structure the survey, we first catalogued the full range of CLARIN.SI's existing tools and services, and organized the sections around key anticipated user activities, such as depositing or downloading resources, using tools, and attending events. An additional section addressed potential future activities, focusing on areas not yet covered by the infrastructure. To allow for a better understanding of the structure and rationale behind the survey, the complete questionnaire and anonymized survey data are provided at Arhar Holdt et al. (2026).

The survey began by collecting basic participant information and identifying which CLARIN.SI tools and services, if any, they had experience with. Based on their responses, participants were directed to relevant subsections where they could evaluate specific aspects of the infrastructure. Each section also included an open-ended question that invited participants to provide more information on what they felt was missing, allowing for broader, unanticipated input.

The Slovenian survey ran from 28 February to 1 April 2025, while its English counterpart was open from 10 November to 10 December 2025. The surveys were disseminated through academic, educational, and professional networks in Slovenia and Europe judged to have members who are current or

potential CLARIN.SI users. According to the 1KA platform statistics, 895 individuals entered the survey introduction, 296 proceeded to the first page, 235 started responding, 228 submitted at least a partial response, and 169 completed the questionnaire. We therefore analyse 228 valid partial or complete responses. The majority of these are from the Slovenian version (170 respondents), and 58 (25%) from the English version of the survey. The average completion time was 11 minutes and 2 seconds. Although the survey included prompts encouraging participants to answer all questions, they were allowed to skip items, so the number of responses (denoted below by N) varies by question.

The survey data were analysed using descriptive statistics, with a focus on aggregate distributions rather than inferential comparisons. This reflects the exploratory aim of the study, namely to map user profiles, experiences, and development priorities relevant to CLARIN.SI rather than to test predefined hypotheses. The findings below are therefore presented holistically, with occasional identification of differences between the Slovenian- and English-language surveys; these differences should be interpreted as indicative rather than statistically confirmed (see also Limitations). For the open-ended questions, we applied content analysis using a bottom-up, inductive approach (Krippendorff, 2018). Responses were coded using generalized words or phrases reflecting the key concerns, observations, and development priorities identified by participants. These results are summarised in the corresponding sections.

3 Results

3.1 The Participants

The affiliations (N=226) of participants was mainly with academic and research institutions (84%). 43% worked or studied at a higher education institution, and 41% reported working in a research institution. Some were employed in private or public companies (6%), public administration (3%), primary or secondary schools (2%), cultural institutions (1%), and non-profit organizations (1%). An additional 2% selected “Other,” for instance, media houses and libraries. Interestingly, the distribution of affiliations differs significantly between Slovenian and international survey participants. Slovenian participants are primarily affiliated with research institutions (47%), followed by higher education institutions (35%). In contrast, most international participants come from higher education institutions (67%), with fewer from research institutions (24%). The results also point at a more diverse profile of Slovenian users in comparison to foreign users.

In terms of professional roles (N=226), over half of the respondents (54%) were engaged in research, including doctoral and postdoctoral work. Others were students (16%) or educators (11%) at different education levels, or involved with technical and development work (8%), administrative tasks (4%), and management (3%). The remaining 4% specified other roles, such as translation or language editing. Disciplinary backgrounds (N=227) were mainly in the humanities (59%), engineering and technology (18%), social sciences (15%), followed by natural sciences (7%), and medicine (1%).

Most participants (N=225) reported they conduct language-related work, however, 10% reported no language-related activities. Their main fields of work (multiple answers possible) were language technologies (38%), linguistic description (36%), natural language processing (28%), translation (20%), theoretical linguistics (20%), and language teaching (19%). Other fields, such as language standardization (13%), history (5%), literary studies (4%), journalism (4%), sociology (2%), library and information science (2%), law and criminology (2%) and political science (1%), were also represented. More Slovenian than international survey participants are active in language standardization (14% vs. 9%). In contrast, international participants are significantly more involved in linguistics: 35% are active in theoretical linguistics (compared to 14% in Slovenia), 55% in language description (30% in Slovenia), and 28% in language teaching (16% in Slovenia). Additionally, 5% of international participants come from political science, a field not represented among Slovenian respondents.

Figure 1 presents the distribution of disciplinary backgrounds among survey participants working in the fields of natural language processing (NLP) and language technologies (LT). Contrary to the common assumption that these areas are primarily rooted in machine learning and therefore dominated by engineering and computer science, the survey results reveal a strong presence of researchers from the humanities. In the case of natural language processing, the share of participants with a humanities back-

ground is equal to that of participants from engineering and technology disciplines. For language technologies, this trend is even more pronounced, with more than half of the survey participants identifying a humanities background. Overall, the figure highlights the strongly interdisciplinary nature of both fields and underscores the substantial role that researchers with a background in humanities play within the CLARIN.SI user community.

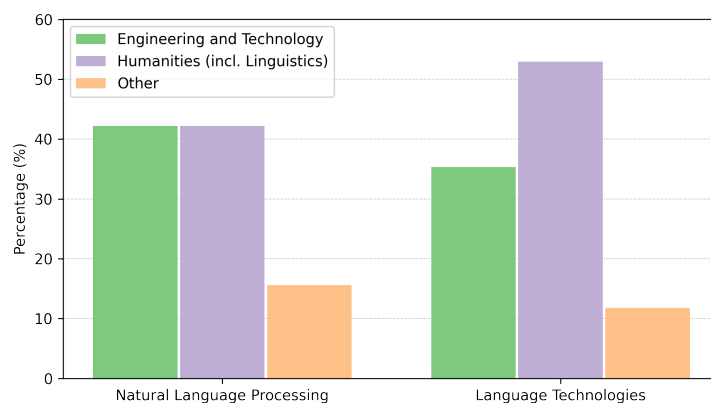


Figure 1: The distribution of disciplinary backgrounds among survey participants working in the fields of natural language processing (NLP) and language technologies (LT).

The participants (N=227) were diverse in age, with the largest groups aged 45–54 (28%), 35–44 (25%), and 25–34 (21%). Smaller segments included respondents aged 55–64 (13%), 18–24 (9%), 65 or older (4%), and two respondents under 18. 60% of the participants were women, 34% were men, 3% another gender, and 3% preferred not to disclose.

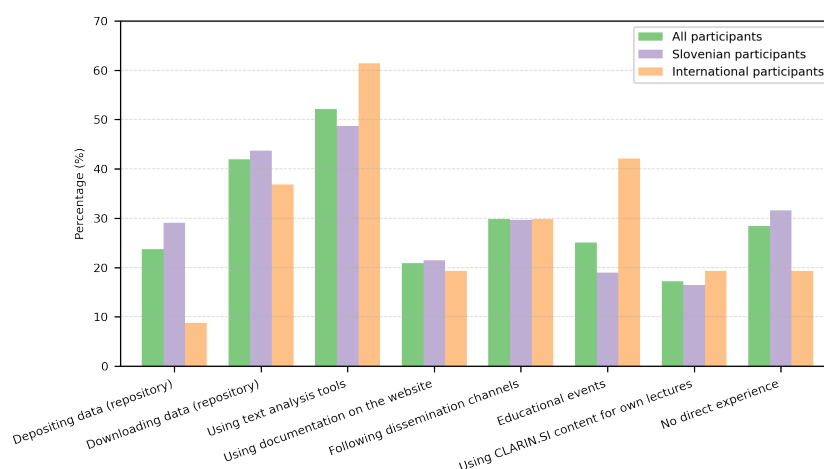


Figure 2: Reported experience of survey participants with the CLARIN.SI activities and services (multiple responses allowed).

The participants (N=215) reported their experience with various components of the CLARIN.SI infrastructure (multiple answers were possible). As shown in Figure 2, 24% had deposited resources, and 42% had downloaded resources from the CLARIN.SI repository. 52% had experience with the CLARIN.SI concordancers and text analysis tools. 21% used documentation on the CLARIN.SI website, 30% followed CLARIN.SI news through website or social media, 25% had attended training or events about CLARIN.SI, and 17% had integrated CLARIN.SI into their own teaching or presentations. Nearly one-third of respondents (28%) reported no direct experience with the infrastructure.

When comparing Slovenian and international participants, the former have more often deposited resources in the CLARIN.SI repository (29% vs. 9%), as well as downloaded and used data from the

repository (44% vs. 37%). In contrast, a greater share of international participants used the CLARIN.SI concordancers and text analysis tools (61% vs. 49%) and have attended training sessions or events promoted on the CLARIN.SI website (42% vs. 19%). These differences may reflect the influence of the CLASSLA-Express international workshop series (Ljubešić et al., 2024), which, while it included a few events in Slovenia, was mostly held in neighbouring countries.

3.2 Depositing Resources in CLARIN.SI

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 and now archives around 700 entries.

The CLARIN.SI repository is widely used by the survey participants – a quarter of them had already deposited their own resources there, and almost half of them have downloaded and used resources from the repository. Among respondents who reported depositing resources in the CLARIN.SI repository, those who completed the follow-up questions (N=51) reported an overall positive experience. The highest-rated items on a 5-point scale included trust in data storage (average with standard deviation: 4.69 ± 0.87), the importance of deposition for open science (4.76 ± 0.85), the belief that deposition of resources supports their reuse in the community (4.41 ± 1.08) and increases resource visibility (4.21 ± 1.17). The efficiency of the editorial review process was rated highly (4.49 ± 0.92), but the ease of different technical aspects slightly lower: the ease of license selection (4.04 ± 0.95), submission workflow (3.87 ± 0.92), technical file preparation (3.61 ± 1.14), and metadata entry (3.67 ± 0.87).

The majority of respondents who deposit resources in the repository (63%) reported consulting CLARIN.SI support, including website instructions and helpdesk assistance. Of these respondents, 30 completed the follow-up questions, and they rated the helpdesk highly: communication clarity and friendliness averaged 4.45 ± 1.09 , and effectiveness and timeliness of help averaged 4.62 ± 1.05 . The availability (4.0 ± 1.0) and clarity (4.04 ± 1.02) of instructions were rated slightly lower.

When asking for optional open-ended comments, we received 33 valid responses, 17 of which provided concrete suggestions.¹ Consistent with the quantitative ratings above, several thematic clusters were identified. The largest indicated need for simpler depositing process, including intuitive and automated deposition workflow and less demanding editorial control. A second cluster concerned repository entry management, highlighting simplified versioning, clearer metadata management, automated processing of large datasets, and practical support materials (e.g. video tutorials). The final cluster related to post-deposit visibility, with calls for better resource promotion and search functionality, as well as enhanced resource monitoring and authorship visibility.

3.3 Finding and Downloading Resources from CLARIN.SI

Almost half of the survey participants (N=90; 42%) reported using the CLARIN.SI repository to find and download language resources. The subset of users who responded to the follow-up questions (N=81) reported an overall positive experience. The most highly rated aspects were download reliability and speed (avg. with std = 4.61 ± 0.68), followed by diversity (4.36 ± 0.75) and the quality (4.28 ± 0.86) of available resources. Participants also emphasized that CLARIN.SI resources are important for their work (4.11 ± 0.97) and help save time in preparing their own data (4.28 ± 0.85). Licensing information was considered clear and understandable (4.25 ± 0.8). Slightly lower, but still positive ratings were given to the recency of resources (4.16 ± 0.85) and the suitability of formats for further use (4.13 ± 0.88). The lowest scores were given to the metadata clarity (3.97 ± 0.74), and the ease of search (3.87 ± 0.99).

Of the 45 valid open-ended responses in this section of the survey, 28 included at least one improvement suggestion. In line with the quantitative ratings above, several thematic clusters were identified, which relate to specific functionalities of the repository: a) search functionalities and downloads (e.g. clearer filtering, better indexing in external search engines, more visible download options); b) more consistent metadata management (easier versioning, addition of Slovenian metadata translations); c) resources and formats (stronger resource quality control, broader format support, more lightweight formats,

¹ In all open-ended questions, some participants indicated they had no specific suggestions, e.g. "I", "Nothing to add", "Keep up the good work", etc.

and more resources overall); and d) user help (inclusion of user ratings and forums, simpler solutions for non-experts, and search video tutorials).

3.4 Using CLARIN.SI Concordancers and Language Tools

CLARIN.SI hosts several concordancers, in particular noSketch Engine (Kilgariff et al., 2014) and KonText (Machálek, 2020). Furthermore, we have three installation of noSketch Engine: the old “Bonito”² and the new “Crystal” version, as well as Crystal that requires log-in but allows more functions. Two further Slovenian-only on-line tools are included among the CLARIN.SI services: Korpusnik (Kosem et al., 2023), gives an overview of word usage based on querying five Slovenian corpora through the concordancer, while SENTA aims to simplify complex Slovenian sentences while preserving the original meaning (Žagar et al., 2024).

About half (52%) of survey participants reported using the CLARIN.SI concordancers and text analysis tools. Among those who responded to the follow-up questions (N=88), levels of user engagement varied across tools. Crystal (57%) and Bonito (46%) are the most widely used, followed by the KonText concordancer (33%), and NoSketch Engine with login (29%). Slovenian-only tools Korpusnik and Senta are used by 35% and 16% of Slovenian survey participants, respectively.

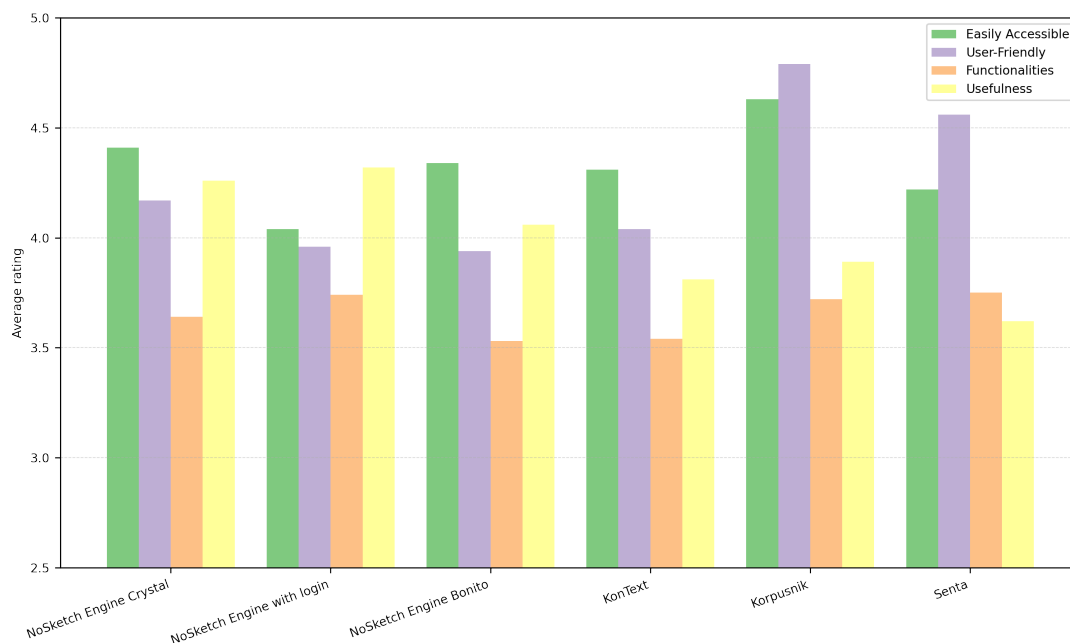


Figure 3: Average ratings of the CLARIN.SI text analysis tools in terms of their accessibility, user-friendliness, provided functionalities, and their usefulness for the participants’ work.

Figure 3 compares the CLARIN.SI concordancers and other text analysis tools in terms of accessibility, user-friendliness, available functionalities, and perceived usefulness for participants’ work. All tools were rated as easily accessible, with Korpusnik (avg. with std = 4.63 ± 0.60) and NoSketch Engine Crystal (4.41 ± 0.91) rated the highest, and the NoSketch Engine concordancer with login rated the lowest (4.04 ± 1.15). User-friendliness varied notably: the Slovenian tools Korpusnik (4.79 ± 0.42) and Senta (4.56 ± 0.53) were considered significantly more user-friendly than the concordancers, with NoSketch Engine Crystal (4.17 ± 0.93) rated the most user-friendly among the latter. According to participants, the NoSketch Engine concordancer with login offers the most functionalities (3.74 ± 1.10), while NoSketch Engine Bonito (3.53 ± 0.98) and KonText (3.54 ± 1.10) offer the least. In terms of overall usefulness, the NoSketch Engine concordancer with login (4.32 ± 0.99) and NoSketch Engine Crystal (4.26 ± 1.03) were rated as the most helpful for the participants’ work.

²Note, however, that we have recently shut down Bonito, as it was continuously and very aggressively harvested by robots, leading to outages of the infrastructure.

Although improving these third-party tools is outside our scope, we still collected open-ended user suggestions to gain better insight into perceived weaknesses of each tool. Among the concordancers, Crystal received the most detailed feedback (29 responses; 18 with suggestions), with users highlighting usability and learnability issues, especially the unintuitive interface with functions hidden behind icons. Bonito (21 responses; 10 with suggestions) drew mixed feedback, with some valuing its minimalist interface and stable features, while others noted concerns about maintenance, accessibility, and outdated interface elements. NoSketch Engine with login (11 responses; 6 with suggestions) received modest feedback focused on clearer settings, better onboarding for non-technical users, and alignment with Sketch Engine functionality. Regarding KonText (16 responses; 4 with suggestions), some users noted the interface’s complexity and a lack of tooltips and reminders on how to use CQL query language. On the other hand, for Korpusnik (13 responses; 3 with suggestions), comments pointed to the need for more flexible display options and support for multiword lexeme search, while feedback on Senta (5 responses; 3 with suggestions) focused on improving summarisation quality and extending the tool to spoken-language inputs with LLM-supported explanations.

3.5 Staying Informed about CLARIN.SI

A third of survey participants regularly or occasionally follow CLARIN.SI news through its website or social media channels. Among respondents who completed the follow-up questions focused on communication and updates from CLARIN.SI (N=51), most reported receiving information via the CLARIN.SI website (54%) and the CLASSLA mailing list (43%), while fewer used channels such as LinkedIn (26%), X/Twitter (18%), and Discord (14%). The distribution of Slovenian and international users across these channels varies notably. Nearly twice as many international participants follow the CLASSLA mailing list (64%) compared to Slovenian participants (35%), while the opposite is true for news on the CLARIN.SI website, which is followed by 62% of Slovenian and 36% of international users. The proportions of users following CLARIN.SI on X and Discord are similar across both groups, but significantly more international participants follow CLARIN.SI on LinkedIn (43% vs. 19% of Slovenian participants).

Participants found the news clear and engaging (4.13 ± 0.74), and the frequency of updates generally appropriate (4.09 ± 0.82). There was less agreement with statements that the updates motivated them to use CLARIN.SI services or resources (3.85 ± 1.02), that they were useful for their work (3.83 ± 0.99), and that they were sufficiently visible (3.44 ± 1.09). The idea of a dedicated CLARIN.SI newsletter received mixed support (3.85 ± 1.09).

In this section, we received 27 valid open-ended responses, 8 of which included improvement suggestions. Several called for wider dissemination and promotion, including a newsletter (2 participants), stronger promotion of resources (3), especially new/updated resources (3), guidance on using tools and resources (2), plus more visible information about events (1) and a communication style that better serves both new and experienced users (1) and comprises shorter, more focused messages (1).

3.6 Attending CLARIN.SI Related Presentations and Events

A quarter of all survey participants reported attending CLARIN.SI training or events, including lectures, workshops, and other activities. Among respondents who completed the follow-up questions (N=39), most referenced participation in the Language Technologies and Digital Humanities conferences, held biennially in Ljubljana, and related events, including the JOTA lectures and the CLASSLA-Express (Ljubešić et al., 2024) workshops.

The overall experience with CLARIN.SI events (N=40) was rated highly: the average rating of usefulness for their work or research was 4.41 ± 0.68 , with even higher scores for clarity (4.76 ± 0.43), appropriate level of expertise (4.82 ± 0.39), and event organization (4.67 ± 0.53). Participants also appreciated the opportunities for interaction (4.64 ± 0.63) and the accessibility of materials after the event (4.75 ± 0.6).

Out of 22 valid open-ended responses in this section, 10 included at least one improvement suggestion. The most common requests were for more events focused on practical usage and usability of tools/resources (3 participants), longer events to allow more time and tasks (2), and better promotion/communication about events (2); two participants also wanted a wider thematic scope (e.g., more

digital humanities, beyond language technologies). In addition, two participants asked for more content on corpus data analyses (including methods/statistics and corpus-building), while single participants mentioned events for community building/networking, sharing good-practice examples, supporting crowdsourcing-oriented resource building initiatives, and offering separate tracks for beginners vs. advanced users.

3.7 Including CLARIN.SI in Presentations and Lectures

Incorporating CLARIN.SI into participants' lectures or presentations (N=37, 17%) was done in various ways (multiple responses were possible). The most common type of integration included citing resources from the CLARIN.SI repository (76%), demonstrating the usability of tools or data through their own practical use (69%), and showcasing tools or resources through live demonstrations (59%). Around half of them referenced CLARIN.SI in discussions of their research methodology (55%) or explicitly directed audiences to the CLARIN.SI infrastructure for further research or learning (52%). A smaller proportion of survey participants incorporated CLARIN.SI into their presentations when explaining concepts such as open science or language resources (38%) or provided comprehensive overviews of CLARIN.SI's structure and purpose (17%; 24% of Slovenian survey participants).

The participants who answered follow-up questions on this topic (N=29) believe that including the CLARIN.SI infrastructure in their lectures or presentations encouraged the attendees to use these resources or tools (avg. with std = 4.05 ± 0.97). However, they did not find CLARIN.SI content to be particularly well-suited or easily accessible for use in presentations. The availability of materials for citation (3.96 ± 1.12), the usefulness of tools and resources for clear demonstration (3.92 ± 0.91), and the accessibility of CLARIN.SI presentation materials providing general information about the infrastructure (3.83 ± 0.98) were rated as less satisfactory than other CLARIN.SI services.

Out of 16 valid open-ended responses in this section, 7 included improvement suggestions, most often asking for ready-to-use presentation materials (3 participants), such as pre-made slides or short "bullet-point" handouts that can be easily reused in talks. Related to this, one participant asked for better access to visual identity elements (e.g., a clearly findable high-resolution logo), while other single-participant suggestions covered better usage-oriented resource presentations with examples, materials tailored for school use/teaching, improved language accessibility for non-experts (Slovenian aside from English), and easier general access to information (through a more appealing UI/more precise explanations).

3.8 Learning about CLARIN.SI

Almost a third of respondents initially stated that they had not (yet) directly used CLARIN.SI in their work. In response to the question "What would encourage you to use the CLARIN.SI infrastructure?", they identified several key motivators, as shown in Figure 4, including better awareness of the infrastructure's structure and purpose (61% of respondents) and more examples of good practice and concrete applications of resources (49%). Many desired more accessible information about updates, events, and opportunities (27%), as well as targeted training or workshops (27%). Less common were requests for increased technical support, a wider range of language resources, and customisation options (10-18%). 12% (6 respondents) indicated that CLARIN.SI was in no context relevant to their work.

In this section, we also received 24 valid open-ended responses, 11 of which included at least one improvement suggestion. Participants' comments were dominated by the desire for better outreach (7 participants) — more promotion and advertising (including social media), clearer introductory information about what CLARIN.SI is, and more information aimed at a wider professional public (including better information about workshops). A smaller set asked for a clearer website (2 participants), especially clearer navigation, and one participant suggested integration of CLARIN.SI-related information into formal education.

3.9 Priorities for Further Development

The respondents (N=121) were asked to rate the importance of various future development priorities for CLARIN.SI on a scale from 1 (not a priority) to 5 (top priority). As shown in Figure 5, the highest-rated priorities were the development of new language tools (avg. with std = 4.06 ± 1.03), integration

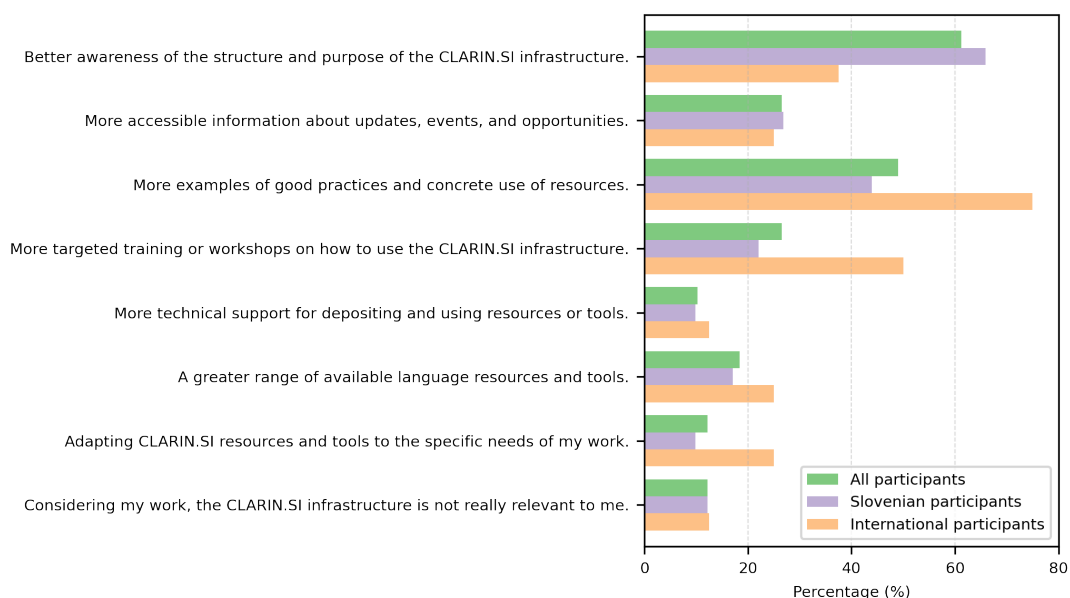


Figure 4: Responses from participants without prior direct experience with CLARIN.SI to the question: “What would encourage you to use the CLARIN.SI infrastructure?” (multiple responses allowed).

of CLARIN.SI with other services (3.94 ± 1.04), and better maintenance and updates of existing tools (3.82 ± 1.1). Improving data search and download processes (3.78 ± 1.15) and increasing the diversity of language resources (3.78 ± 1.16) followed closely.

Slightly lower priorities included improved data quality control (3.71 ± 1.1), increased and targeted training opportunities (3.7 ± 1.18), improved user support and documentation (3.56 ± 1.11), and more frequent and tailored updates on news and events (3.47 ± 1.17). Lower still were clearer presentation of the infrastructure’s composition and accessibility (3.41 ± 1.2), improving the data deposition and storage process (3.4 ± 1.24), and better presentation of CLARIN.SI organization (3.09 ± 1.33). Importantly, a relatively high proportion of respondents selected “I cannot assess” for several items (17–42% respondents), suggesting a need for further clarification of certain aspects of the infrastructure.

In this section of the survey, two open-ended questions were provided. For the first, “If you wish, you can further explain your choices in the previous question,” we received 56 valid answers; 22 contained substantive explanations. In terms of content, respondents most often called for stronger communication and onboarding (4 participants), with clearer basic explanations of what CLARIN.SI offers and how it benefits users, and for more systematic training and education (3) — especially for students and newcomers — covering not only tool use but also methods and examples of good research practices. A second major thread concerned resources and discoverability: four participants stressed expanding and improving corpora and other resources (including filling specific gaps such as 20th-century coverage), while three participants emphasised need for improved repository search capabilities; this was also linked to desires for better web visibility and improved search functionalities. Additionally, suggestions included more regular tool maintenance and updates (3 participants)³, more concrete demonstrations of how resources can be used in practice (2), and several single-participant points: providing clearer information on resource currency, improving integration with other platforms and workflows, preparing reusable presentation materials, clearer site structure, clearer information on how to join the consortium, better quality control of content (not just of the technical aspect), and better corpus annotation.

The second open-ended question, “Is there an area or challenge that we haven’t addressed in the

³For this and similar suggestions, it is not possible to determine whether such observations reflect perceived lower quality, shortcomings in the infrastructure offering, or general observations. This comment, for example, was interpreted as indicating a need for more frequent updates and maintenance of existing tools. All user comments were categorised, although some are more abstract and less suitable for priority development than others.

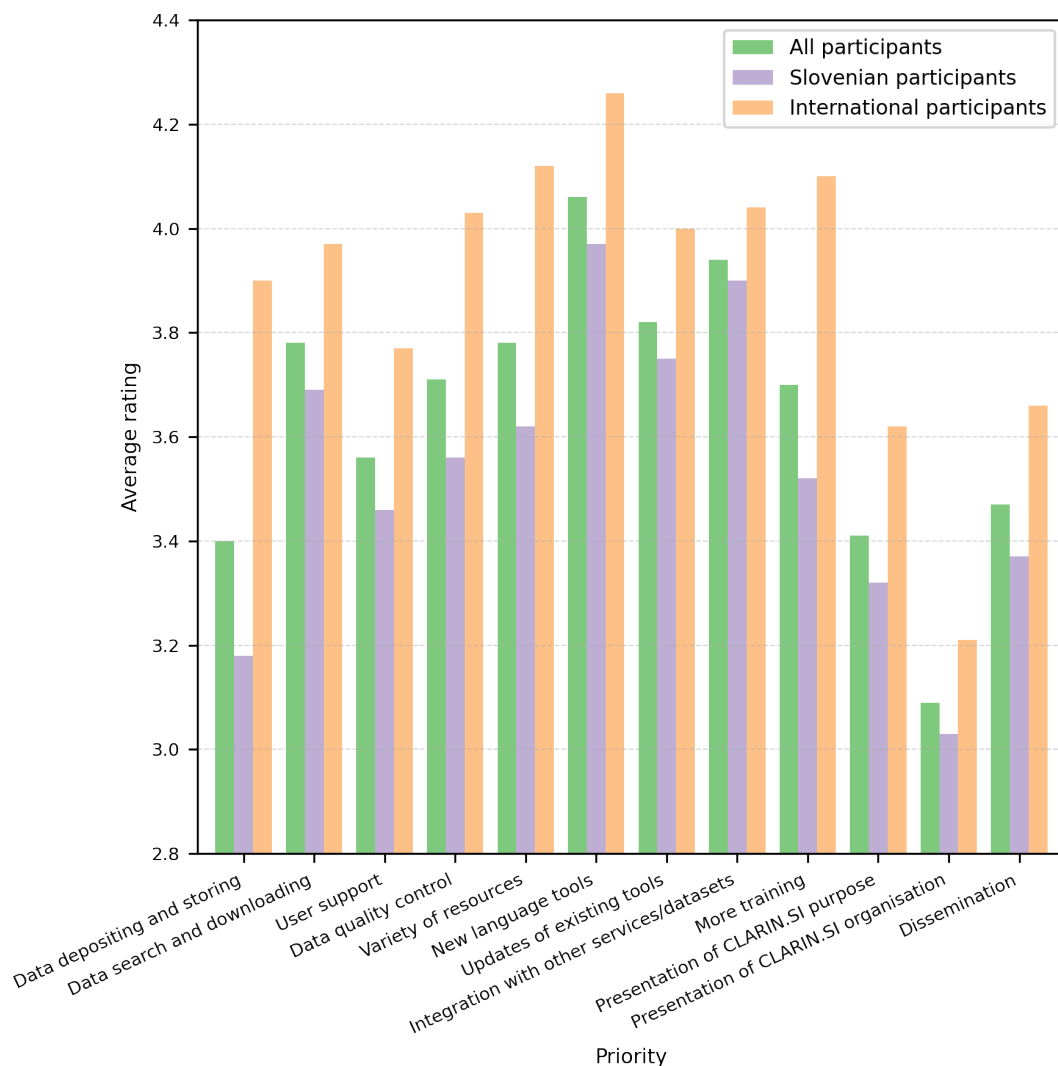


Figure 5: Responses from participants to the question: “How high of a priority for future development are the activities below?” (multiple responses allowed, where 1 represents lowest priority and 5 highest priority).

questions, but which you consider important for the future of CLARIN.SI?” provided 55 valid answers; 16 included at least one substantive comment. The most recurrent themes were better integration with other platforms (2 participants, e.g., Hugging Face and evaluation sites like SloBench), clearer information on consortium membership/participation (2), further development of new resources (2, especially non-standard and spoken/dialect corpora), and more training (2). Other priorities were mentioned by one participant each: clearer guidance on data ownership and personal data protection; better linking of resources across languages; improving visibility/communication to the wider public; improving the concordancers; increasing the value of CLARIN.SI outputs in research evaluation/CVs; more content about AI; clearer quality-control expectations for deposited resources; research support; and improved presentation of available resources and tools that can support educational contexts. As shown, most of these comments align with areas and challenges already addressed in the survey, suggesting that the design effectively anticipated users’ main types of engagement and potential concerns.

4 Conclusion

The survey results provide a comprehensive insight into the types of our users, their experiences, needs, and expectations regarding CLARIN.SI. First, almost a quarter of our users are not NLP specialists nor

linguists but come from other fields of the humanities. While we already have a long-standing collaboration with the Slovenian DARIAH, this indicates that collaboration of the two research infrastructure should be further strengthened to improve our visibility and use among historians, political scientists etc., even more so as we offer a rich set of resources for that could be of use to them, e.g. the ParlaMint corpora of parliamentary debates (Erjavec et al., 2024) and related resources. Such collaboration should also be reciprocal, as DARIAH(-SI) also offers resources and services to support users from primarily linguistic fields.

As far as the experiences of the users go, infrastructure is, overall, well-regarded by its users, particularly in terms of data reliability, open science support, and perceived usefulness of deposited resources. The repository is widely used by participants and provides — overall — a positive experience both with deposition and resource retrieval. Users also value the availability of concordancers and tools that support their work, and were rated highly for accessibility. The communication channels are considered clear and engaging for disseminating information about CLARIN.SI and related activities, although Slovenian and international participants use quite different channels. CLARIN.SI-related events received high ratings for professional quality, clarity, expertise, and the accessibility of materials after the event.

The aim of the survey was to gather feedback from existing users as well as potential new users, with almost a third of participants reporting no previous experience with CLARIN.SI. They highlighted the need to raise awareness of the infrastructure's existence, improve access to information about its structure, and provide more concrete examples of resource use. In contrast, participants with previous experience with the infrastructure identified several priorities for further development. The highest-rated priorities concern the services and tools offered, with the development of new language tools considered the highest priority, followed closely by the integration of CLARIN.SI with other services and by improving existing tools.

Responses to the open-ended questions provided further details on specific issues and participant suggestions. The first major cluster, which includes the most frequent and specific requests, concerns the structure and operation of the CLARIN.SI repository. Responses highlighted search-related issues, such as irrelevant results, limited browsing support, and missing filters, as well as the need to optimise the data deposition workflow (e.g., simplifying submission, editorial review, and entry updates). Open-ended responses included concrete suggestions to address these shortcomings, such as search improvements (autocomplete and quick data previews) and deposition enhancements (clearer, centrally visible instructions and reduced metadata requirements), among others. The technical suggestions are mostly out of our control, as we use the LINDAT DSpace repository software, i.e. it would be difficult to improve its functionality at our end. On the other hand, we plan to install a new version of the repository platform shortly, which may address some of these problems. However, we will take on board suggestion related to documentation, where we should improve its accessibility. But while we understand that depositors would like less work with the metadata entry, we will, however, still insist on clear, complete and harmonised metadata for the repository entries.

In parallel, users repeatedly indicate that CLARIN.SI's impact depends on outreach, onboarding, training, and reusable communication assets, not solely on infrastructure. Suggestions to improve outreach and communication include options such as more concise, action-oriented communications and the preparation of a "CLARIN.SI presentation kit" containing ready-to-use materials. For events, suggestions include more hands-on workshops for different skill levels, longer multi-day formats, and training that goes beyond tool use to cover data processing and research workflows.

Lastly, a separate but still prominent theme identified concerns expanding and modernising the resource offer and formats. Suggestions specifically mention the need for more diverse corpora, including non-standard language and spoken corpora, addressing gaps in diachronic coverage (notably the 1918–1990 period), as well as providing resources in additional formats to support downstream work. However, CLARIN.SI does not, as a rule, produce corpora, and offered resources and corpora are mostly third party resources, so its ability to fill the gaps in the corpora offered is limited. On the other hand, the suggestion to expand the range of deposited formats (in particular, to JSON) could, in certain cases, be acted upon.

For certain suggestions it should also be noted that the conclusions of the survey are somewhat contradictory. For example, it was requested to add the option to monitor downloads of deposited resources, however, this option already exists. Similarly, a simpler approaches to updating resources with new versions was suggested, even though it is already very simple. Such suggestions show that users do not always read the on-line documentation, which is a known and persistent problem, as our reviewing of deposited entries often shows. This again shows that the documentation should be made more prominent and accessible.

The survey results highlighted key areas for improvement, which are also aligned with some specific goals in CLARIN.SI strategy 2024-2030 priorities. The first priority identified in the survey concerns providing more corpus analysis tools, which aligns with the second Pillar of the Strategy (Our offer), specifically *Goal 3.2: Improve the coverage and quality of data, models, and tools offered through the CLARIN.SI infrastructure*. The survey results indicate that there is room for improvement in the provision of high-quality language data sets. Additionally, some comments highlighted the need for better authorship visibility, which is addressed in *Strategy 3.2.6: Act to improve existing citation guidelines and practices in Slovenia*.

More importantly, the survey identified the need to focus on user engagement, which is directly linked to *Pillar 1 (Our users)*, addressing both academic and non-academic user bases. The need for more tutorials and presentations of the infrastructure is also linked to *Strategy 1.1.2 (Improve the communication channels to better reflect the mission, network, and offer of CLARIN.SI)* and *Strategy 1.1.3 (Improve services through newly created training, workshops, and online tutorials)*. The highlighted priorities are relevant to the key activities of CLARIN nodes research infrastructures, as well as research infrastructures in general, i.e. user base and infrastructure offering. For CLARIN.SI the survey results and actionable user-identified priorities will serve as guidance for further development.

5 Limitations

The study relies on descriptive statistics, in line with its exploratory aim of collecting structured feedback on the quality of CLARIN.SI services, tools, and resources, and of identifying areas for improvement relevant to future development and strategic planning. No predefined hypotheses were posed, and the analysis was therefore not designed as a basis for inferential testing. Inferential comparisons between respondent subgroups, such as Slovenian and international participants, were consequently not conducted. This was also due to the structure of the questionnaire, which used branching and allowed item non-response, resulting in varying and, for some questions, relatively small numbers of responses, thereby limiting the robustness of such comparisons. The differences observed between groups should therefore be understood as indicative rather than statistically confirmed. A further limitation concerns representativeness. Participation was voluntary, and the survey was distributed through academic, educational, and professional networks considered likely to reach existing or potential CLARIN.SI users. The sample cannot be considered representative of any broader population, and the findings may also reflect self-selection bias, including the possible overrepresentation of respondents who are more aware of the infrastructure or more motivated to provide feedback. At the same time, the inclusion of participants without prior direct experience with CLARIN.SI broadens the range of perspectives captured by the survey and offers insight into factors that may affect future uptake. These limitations should be taken into account when interpreting the findings.

6 Acknowledgements

The authors would like to thank the anonymous reviewers for their useful comments. The work was supported by the Slovenian Research and Innovation Agency (ARIS) via the research infrastructure CLARIN.SI (I0-E004) and the core research programmes Language Resources and Technologies for Slovene (P6-0411) and Knowledge Technologies (P2-0103).

References

- Arhar Holdt, Š., Meden, K., Kuzman Pungersšek, T., Ljubešić, N., & Erjavec, T. (2026, February). CLARIN.SI user survey 2025. <https://doi.org/10.5281/zenodo.18486462>
- De Smedt, K. (2020). *Clarino - user survey report* (tech. rep.). CLARINO. <https://clarin.w.uib.no/files/2020/08/CLARINO-User-Survey-Report.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, Starkadur, 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., Ljubešić, N., Meden, K., Kuzman, T., Laskowski, C., Javoršek, J. J., Krek, S., Jemec Tomazin, M., Dobrovoljc, K., Holdt, Š. A., Lenardič, J., & Fišer, D. (2025). CLARIN.SI, the Slovenian node of CLARIN: Ten years on. *Selected papers from the CLARIN Annual Conference 2024, Linköping Electronic Conference Proceedings (ECP)*. <https://doi.org/10.3384/ecp216.07>
- 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., 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
- Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.
- Ljubešić, N., Kuzman, T., Filipović Petrović, I., Parizoska, J., & Osenova, P. (2024). CLASSLA-Express: a Train of CLARIN. SI Workshops on Language Resources and Tools with Easily Expanding Route. *CLARIN Annual Conference Proceedings 2024*, 31–35.
- Luzietti, R. B., Spadi, A., Giampietro, N., Mancuso, G., Caravale, A., D'Eredità, A., Caradonna, M., Moscati, P., Quochi, V., Monachini, M., et al. (2025). Digital humanities and heritage science: Moving from landscaping to a dynamic research observatory in an open science cloud. *Umanistica Digitale*, (20), 419–439.
- 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>
- Monachini, M., Quochi, V., Luzietti, R. B., Spadi, A., Mancuso, G., D'Eredità, A., Caravale, A., Giampietro, N., Caradonna, M., & Melaccio, D. (2025, October). D2.2 updated report on the h2iosc landscapes. <https://doi.org/10.5281/zenodo.17737101>
- Nicolas, L., König, A., Monachini, M., Del Gratta, R., Calamai, S., Abel, A., Enea, A., Biliotti, F., Quochi, V., & Stella, F. V. (2018). CLARIN-IT: State of affairs, challenges and opportunities. *Selected papers from the CLARIN Annual Conference 2017*. <https://hdl.handle.net/10863/8194>
- Sichera, P., Monachini, M., Quochi, V., Giampietro, N., Fabiani, V., Ottaviani, R., & Luzietti, R. B. (2025). Synergies between clarin-it and operas-it within h2iosc: Monitoring communities and orchestrating digital services. *CLARIN Annual Conference Proceedings*, 26.
- 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://aanthology.org/2024.lrec-main.1279/>