

Govorjena slovenščina: A Platform for the Structured Collection of Conversational Speech

Darinka Verdonik, Andreja Bizjak, Gregor Donaj

University of Maribor, Slovenia

{darinka.verdonik, andreja.bizjak1, gregor.donaj}@um.si

Abstract

Conversational spoken language resources remain underrepresented despite their importance for linguistic research and speech technology development. Their collection is constrained by methodological, technical, and legal challenges, particularly in the case of private, everyday interaction. This paper presents Govorjena slovenščina, a CLARIN.SI service developed to support the systematic collection, transcription, and archiving of spoken Slovene. The platform integrates standardized workflows for consent management, metadata collection, data submission, and transcription, and is designed as a reusable research infrastructure rather than a project-specific collection tool. Drawing on insights from citizen-science and crowdsourcing initiatives, the paper discusses key design choices related to participant motivation, data quality, and legal governance, and reports on initial corpora contributions enabled by the platform. The approach demonstrates how a standards-driven, legally grounded infrastructure can support the sustainable development of conversational speech resources, particularly for smaller or under-resourced languages.

1 Introduction

Spoken language resources remain considerably less developed than written ones, especially with regard to data representing everyday human–human interaction. Nevertheless, spoken data are indispensable for achieving comprehensive and empirically grounded linguistic descriptions, as they make it possible to observe variation, spontaneity, and context-sensitive language use that are largely absent from written texts. Analyses of spoken interaction provide crucial theoretical insights into how speakers manage turns, negotiate meaning, coordinate social actions in real time, etc. In addition, spoken corpora play a key role in the advancement of speech technologies, most notably automatic speech recognition and dialogue systems, which rely on authentic spoken input to function robustly in natural communicative settings. The continued scarcity of high-quality spoken resources therefore limits both linguistic theory-building and the development of language technologies capable of handling informal, interactional speech.

Substantial quantities of spoken data can be obtained from readily available sources such as parliamentary proceedings, broadcast media, or online platforms including YouTube. While these resources are valuable, they cannot replace a broad and representative range of private data, as speech practices differ markedly across communicative domains. Previous research has demonstrated that conversational data are in higher demand and used more frequently in linguistic research than context-governed data (Love et al., 2017). Despite this demand, conversational corpora remain scarce as they are among the most difficult resources to compile. Their collection requires considerable organizational effort, including the recruitment and coordination of a wide network of field collaborators and volunteer speakers. Moreover, potential participants may be reluctant to consent to the recording of their speech for personal or privacy-related reasons. The processing and storage of spoken data also present significant challenges, as they necessitate transparent data management policies and strict adherence to personal data protection regulations. Further complications arise from technical factors affecting recording quality, such as background noise, variability in recording equipment, and uncontrolled recording environments. Finally,

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

achieving adequate speaker diversity with respect to age, gender, dialectal background, and sociolectal variation is essential for representativeness, yet remains difficult in practice, further constraining the development of balanced conversational language resources.

This paper presents an online platform Govorjena slovenščina (Spoken Slovene)¹, launched in March 2025 as part of the LLM4DH project and financed by the Slovenian Research and Innovation Agency. The platform is designed to address the challenges of conversational data collection by providing scenarios and guidelines for field recording of conversational data, consent forms for speakers to allow the use of recordings, standardized forms for documenting information about recordings and speakers, remote submission of the data, secure and long-term archiving of the data, and unified guidelines for transcribing speech in Slovene. The platform is offered as the CLARIN.SI service for speech data collection, and is provided by the University of Maribor, Faculty of Electrical Engineering and Computer Science as the CLARIN.SI consortium partner.

This paper focuses on challenges related to creating the Govorjena slovenščina platform, which are: (1) Resolving legal aspects of conversational speech data collection and distribution, (2) Attracting participating individuals, (3) Defining scenarios for who, where, what, and how to record, (4) Standardizing the metadata on speakers and recordings, (5) Facilitating recording submission and data management and (6) Standardizing and supporting the transcription processes digitally.

2 Citizen Science and Crowdsourcing in the Collection of Spoken Language Resources

Traditional laboratory and field-based approaches remain essential for producing high-quality spoken language data, yet they are difficult to scale when the research objective is to document broad social and regional variation in everyday speech. In particular, the collection of spontaneous interaction is inevitably affected by the observer's paradox, as formulated by William Labov (Labov, 1972): speakers tend to adjust their language use once they become aware that they are being recorded. This effect persists regardless of whether recordings are made by researchers, by speakers themselves, or through online interfaces, and thus cannot be fully eliminated by changing the recording setting alone.

In response to these constraints, increasing attention has been paid to remote and large-scale data collection strategies based on citizen science and crowdsourcing. Enabled by widespread access to digital technologies and mobile devices, these approaches make it possible to reach large and diverse speaker populations at comparatively low cost. At the same time, crowdsourcing is not a monolithic method: it encompasses paid microtask platforms, volunteer-based citizen science initiatives, and alternative incentive models such as Games-With-A-Purpose (GWAP) and Collect4NLP. Each of these models differs with respect to sustainability, participant motivation, cost structure, and the degree of control over data quality (Lyding et al., 2022).

Moving speech collection online, however, introduces new design challenges (Eskenazi et al., 2013). The potential to reach many contributors must be balanced against risks related to authenticity, representativeness, and technical quality. Task design is particularly critical: insufficient guidance may result in unusable recordings, while overly detailed instructions can unintentionally shape participants' speech and reduce natural variation. Successful initiatives therefore tend to combine simple, user-friendly interfaces with carefully calibrated instructions, basic quality-control mechanisms, and transparent communication about data use.

To identify effective strategies and transferable design principles, a review of 45 crowdsourcing-based speech collection initiatives was conducted (Bizjak, 2025). On the basis of contributor scale, collection goals, targeted speech types, and other, three European projects were selected for closer analysis: CorCenCC (Welsh), Anneta kõnet (Estonian), and Lahjoita puhetta (Finnish). These projects illustrate how technical solutions, legal and ethical frameworks, and motivation strategies can be combined to support large-scale collection of spoken language data, providing valuable reference points for the design of the Govorjena slovenščina platform.

CorCenCC (2016–2020) represents one of the most comprehensive national corpus-building efforts for a lesser-resourced language, collecting over 11 million words in total, including 2.8 million words of spo-

¹<https://govorjena-slovenscina.um.si/>

ken language, of which approximately 240,000 words were recorded in private, informal settings (Knight et al., 2020). A distinctive feature of CorCenCC is its “bottom-up” corpus design, aligned with the participatory ethos of citizen science. Participation was supported through a dedicated mobile application for iOS and Android, as well as a web interface, enabling contributors to submit recordings together with consent documentation and sociolinguistically relevant metadata (Knight et al., 2001; Neale et al., 2017). The project’s emphasis on detailed speaker profiling supported representativeness and balancing goals, while its governance model foregrounded transparency and participant control over contributions. Notably, the project also faced challenges common to smaller language communities: despite extensive media presence and outreach, private, informal recordings proved difficult to obtain, plausibly due to heightened concerns about identifiability in a relatively small speech community (Knight et al., 2020; Neale et al., 2017). This illustrates an important general lesson for spontaneous speech collection: privacy concerns can be a stronger limiting factor than technical barriers, and trust-building measures (clear anonymization procedures, accessible documentation, and repeated explanation of safeguards) are not optional add-ons but central infrastructure.

Annetta kõnet (2022–2023) adopted a donation-based model and collected over 100 hours of Estonian spontaneous speech. A key insight from this campaign is the value of minimizing friction in participation: contributors could simply select a topic and begin speaking, including the option to speak freely without predefined prompts²³. The campaign also used promotional messaging tailored to different target groups—framing participation for younger speakers through the lens of speech technologies in everyday life, and for older speakers through language preservation and cultural identity⁴. This dual framing aligns with broader citizen-science findings that motivations are heterogeneous and that recruitment strategies must be audience-specific rather than generic. It also demonstrates how “purpose communication”—explaining how data will be used and why it matters—can function as a motivational lever and as a trust mechanism.

Lahjoita puhetta (2020–2021) demonstrates the scalability of large-scale remote collection, gathering more than 4,000 hours of Finnish speech within a few months (Lindén et al., 2022). The campaign relied on both web and mobile channels, but it also made a design decision with methodological consequences: to reduce technical complexity and improve signal quality, it prioritized recordings with a single speaker rather than multi-party conversational interaction. This choice highlights a recurring trade-off: multi-speaker conversation may be more interactionally “natural,” but it is also harder to capture with adequate audio quality and clean channel separation in uncontrolled environments. Lahjoita puhetta therefore pursued naturalness through task prompts that elicited spontaneous responses without encouraging read speech, while simplifying the acoustic capture problem. Notably, nearly 90 percent of the 220,000 recordings ranged from 10 seconds to 3 minutes, with an average duration of 30–60 seconds (Lindén et al., 2022). This distribution offers an important operational lesson (Bizjak, 2025): short, low-effort contributions can accumulate into very large datasets, particularly when the interface is simple and the campaign is strongly visible. At the same time, short clips may limit discourse-analytic uses and may not fully mitigate the observer’s paradox, reinforcing the need to align recording formats with research goals (e.g., ASR training vs. interactional linguistics).

Across these case studies, several cross-cutting challenges emerge that are directly relevant to designing sustainable systems for collecting spontaneous speech. First, motivation must be treated as a design problem rather than a “nice-to-have” consideration: why would someone participate, and why would they return (Rutten et al., 2017)? A strong argument in the literature (Eskenazi et al., 2013) is that if payment is used, it should be fair and transparent; token payments can backfire, reduce trust, and complicate ethical governance. Volunteer-based approaches, in turn, require credible purpose communication and visible evidence that contributions are used and valued, not merely collected (Rutten et al., 2017).

Second, quality control requires multiple layers (Eskenazi et al., 2013). Remote speech collection is vulnerable to trivial failures (silent recordings, wrong microphone settings), environmental noise, device

²<https://www.kratid.ee/en/keeletehnoloogia-projektid>

³<https://laecwador.ee/en/projekt/donate-a-speech-campaign>

⁴<https://laecwador.ee/en/projekt/donate-a-speech-campaign/>

variability, and occasional bad-faith contributions (e.g., bots or “fast earners” in paid settings). Best practice therefore includes: (i) clear, concise recording guidance; (ii) lightweight “pretesting” or sample checks; (iii) systematic validation before or after submission (including random sampling and gold-standard comparisons where applicable); and (iv) metadata capture sufficient for diagnosing and filtering problematic recordings. Technical recommendations—such as using lossless formats, adequate sampling rates, and minimizing microphone distance—are important, but they must be translated into user-friendly instructions that do not overwhelm contributors.

Third, legal and ethical governance is foundational for speech resources because speech is a biometric identifier and thus personal data in the GDPR sense. Projects must establish a lawful basis (often consent, sometimes legitimate interest), provide clear privacy policies, and implement anonymization practices (Bizjak, 2025). They must also anticipate downstream access conditions and licensing constraints: permissive licensing supports reuse (including potential commercial applications), but it requires that participant consent and rights management are robust from the outset. Importantly, the case studies show that legal compliance alone is insufficient: participants must understand the governance model in accessible language, and they must feel confident that their contributions will not expose them or others (Knight et al., 2020; Neale et al., 2017).

Taken together, CorCenCC, Anneta kōnet, and Lahjoita puhetta demonstrate how crowdsourcing framework can address traditional bottlenecks in speech data collection—particularly for under-resourced languages—while also revealing the methodological compromises such systems often require. The most successful initiatives treat the platform not merely as a technical upload interface, but as an integrated socio-technical system: it must balance user experience, motivation, trust, legal safeguards, and quality control, while remaining aligned with the linguistic or technological goals of the resulting corpus. These lessons are directly transferable to the Slovene context, where the long-term challenge is not only to collect “more speech,” but to collect speech that is sufficiently diverse, spontaneous, well-documented, and ethically governed to support both linguistic research and robust speech technologies.

In contrast to the large-scale initiatives discussed above—each supported by national institutions, dedicated funding, and extensive media campaigns—Govorjena slovenščina is conceived as a comparatively lightweight academic infrastructure. The project does not aim to compete with national donation campaigns in terms of volume or visibility, nor does it have resources for mobile application development or sustained public promotion. Instead, its primary focus lies in establishing robust legal frameworks, developing reusable technical infrastructure, maintaining unified standards for metadata and transcription, and exploring intrinsic motivational factors that may support voluntary participation in smaller-scale research-driven data collection. The portal is designed as an open service that can be used by individual researchers or project teams from different institutions who wish to rely on its workflows and guidelines for their own speech collection efforts. The project team draws not only on the insights gained from the literature review presented above, but also on its extensive practical experience in speech data collection, accumulated over more than two decades of work on Slovene speech corpora. This experience includes, in particular, the development of the reference speech corpus GOS (Verdonik et al., 2013) and the ARTUR corpus (Verdonik, Bizjak, et al., 2024) dedicated to the development of automatic speech recognition for Slovene.

Govorjena slovenščina does not provide a dedicated recording application; participants use their own recording devices and freely available recording software, and upload the resulting files through the web interface. This design choice deliberately shifts complexity away from custom software development toward standardized procedures and documentation. To support data diversity and prevent overrepresentation of individual speakers, contributions are limited to a maximum of two hours per speaker. At the same time, the platform prioritizes longer recordings, with a minimum duration of five minutes and a maximum of ninety minutes, in order to allow conversations to unfold naturally, reduce initial self-monitoring effects, and provide sufficient contextual depth for linguistic analyses that extend beyond acoustic or lexical levels, including semantic, pragmatic, and interactional research. Furthermore, the platform is conceived not as a project-specific solution, but as a reusable research infrastructure that can be adopted by individual researchers or research teams across institutions for a wide range of speech data collection purposes,

including dialectological, sociolinguistic, pragmatic, lexical, phonetic, or interactional studies, as well as research on speech development, speech disorders, second language acquisition, human–computer interaction, and similar.

3 Platform Govorjena slovenščina

This section outlines solutions for key challenges in creating the Govorjena slovenščina portal.

3.1 Legal Aspects

Legal experts in personal data protection from APHAIA company ⁵ were engaged to design a comprehensive personal data protection protocol in full compliance with European Union legislation, most notably the General Data Protection Regulation (GDPR). Based on their legal assessment, speech recordings of identifiable individuals were classified as personal data, establishing informed consent from the speaker as the most appropriate and transparent legal basis for data processing. The experts further determined that the processing of users' personal data for the purposes of incentives or rewards must rely on a legal basis separate from that governing the collection and processing of speech recordings themselves. In addition, they concluded that the spontaneous speech recordings collected through the portal are unlikely to qualify as copyrighted works, as they typically lack the level of originality required for copyright protection under applicable law.

These legal principles were implemented within the Govorjena slovenščina portal through a set of clearly defined legal and procedural documents. The legal experts prepared a detailed Privacy Policy Statement ⁶, which transparently informs users about the scope, purpose, storage, and processing of personal data. In parallel, contributors are required to accept a Consent for the Publication of Speech Recordings and Related Personal Data, which explicitly covers the recording, storage, annotation, and research use of the donated speech data. In addition, the Terms of Contribution of Recordings and Contest Participation regulate the conditions under which users may submit recordings and, where applicable, participate in incentive-based activities hosted on the portal. All documents are made accessible directly through the portal interface, ensuring that contributors are fully informed prior to participation. Before publication, all data are anonymised by removing personal identifiers. Speakers' names and surnames are deleted from metadata, ensuring that individuals cannot be easily identified.

3.2 Attracting Participating Individuals

As the Govorjena slovenščina portal lacks the financial resources to support a large-scale marketing campaign, particular emphasis has been placed on identifying and fostering intrinsic motivational factors that encourage sustained participant engagement. The portal homepage therefore foregrounds motivations that appeal to contributors' sense of social value and personal relevance, such as the opportunity to donate speech in order to preserve samples of Slovene and its regional varieties as part of the national linguistic and cultural heritage. Additional emphasis is placed on supporting the development of digital language technologies for Slovene and its dialects, as well as contributing to research on child speech and language development. To complement these intrinsic incentives, the portal also offers modest but tangible rewards, including items such as memory cards or headphones, which are awarded to participants who submit recordings that meet quality criteria and reach a cumulative duration of 50 minutes. This hybrid approach balances voluntary participation with light extrinsic encouragement, while avoiding the pressures associated with monetary compensation.

A second key focus of the portal is the establishment of trust between contributors and the research team, as trust is a prerequisite for individuals to feel comfortable donating personal speech data. The portal therefore provides transparent and accessible information on how recordings will be processed, stored, anonymized, and used for research and technological development. Clear explanations of data protection measures are accompanied by information about the institutions and individuals responsible for maintaining the portal. In addition, the portal publishes popular-science articles and explanatory texts

⁵<https://aphaia.co.uk/>

⁶<https://govorjena-slovenscina.um.si/uporaba-posnetkov/>

about spoken language and speech research, which serve both to disseminate knowledge and to reinforce the public-interest orientation of the project.

Third, the portal places strong emphasis on the clarity of task descriptions in order to minimize uncertainty, which is a known barrier to participation in crowdsourced data collection. Contributors are provided with concise and comprehensible instructions that guide them through the recording process step by step. To further support participation, the portal outlines a range of possible recording scenarios and suggests concrete speech topics, helping contributors understand what kinds of recordings are desired while still allowing for spontaneity and naturalness in speech production. These instructions are complemented by an explanatory video with illustrative example that demonstrates the recording procedure in a clear and accessible manner.

Finally, considerable attention has been devoted to the development of a clear, user-friendly, and visually appealing web interface with an engaging and playful aesthetic. The design incorporates illustrative graphics and examples of authentic speech drawn from previously collected conversational data. In addition, the portal features two introductory videos that serve complementary communicative functions: the first video is designed to attract and motivate potential participants by presenting the initiative in an accessible and engaging manner; the second video explains the importance, intended use, and broader societal and scientific value of the collected speech data, while the third video, as outlined above, provides a step-by-step demonstration of how to create and submit a recording. Together, these visual elements support user engagement, enhance transparency, and contribute to a more intuitive and reassuring user experience.

The usability of the platform has been critically evaluated by experts in linguistics and design using a heuristic evaluation based on Nielsen's usability principles (Nielsen, 1994), systematically assessing interface design, feedback mechanisms, consistency, and error prevention. Based on this evaluation, iterative improvements to both the interface and the task workflow were implemented.

3.3 Recording Scenarios and Metadata Standardization

Conversational data are often categorized in a relatively limited and abstract manner. Existing classification schemes typically distinguish between parameters such as direct versus distanced contact, dialogue versus multilogue, and broad domains including face-to-face conversation, telephone interaction, interviews, or business transactions (Oostdijk, 2000). In the Slovene reference corpus GOS (Verdonik et al., 2013), conversational data were further specified through domains such as conversations among friends or within families. While such typologies are useful for corpus design and analysis, they offer limited practical guidance to contributors engaged in data donation, particularly with respect to what constitutes an appropriate conversational form, level of spontaneity, or choice of topics.

To address this limitation, the Govorjena slovenščina platform places strong emphasis on providing concrete, illustrative guidance for participants, both in terms of speech genres and topic selection. Contributors are encouraged to produce recordings that fall within a set of clearly defined and easily recognizable genres, including interviews, discussions, instructions, narrations, explanations, and informal or casual conversations. These genres are described through short, example-based explanations grounded in everyday communicative situations, using non-technical language and avoiding specialist linguistic terminology. They are selected to capture a broad spectrum of communicative functions, such as information exchange, opinion expression, procedural discourse, and narrative structuring. In addition, the platform proposes a wide range of everyday topics that reflect natural spoken interaction, such as hobbies and leisure activities, sports, music, travel experiences, technology use, food and cooking, health and well-being, environmental issues. Importantly, contributors are explicitly advised to select only topics they feel comfortable sharing publicly, thereby reinforcing awareness of the public and research-oriented nature of the recordings and supporting informed participation.

Special attention is also devoted to technical and procedural guidelines aimed at ensuring adequate recording quality without imposing excessive technical demands on contributors. The platform provides step-by-step instructions on how to prepare for recording, including recommendations to choose a quiet environment, reduce background noise (e.g. television, traffic, or other speakers), and position the record-

ing device at an appropriate distance from the speaker’s mouth. Guidance is given on using common devices such as smartphones or laptops, adjusting basic recording settings, and avoiding handling noise. To further reduce uncertainty, contributors are offered the possibility to make and submit a short test recording before donating longer samples. Together, these measures are designed to balance recording quality with accessibility, enabling contributors with varying levels of technical expertise to participate while producing speech data suitable for linguistic analysis and technological applications.

Participating citizens are required to collect metadata on both speakers and recorded events alongside the recordings. The categorization of these metadata is based on the Slovene reference speech corpus GOS 2.0 (Verdonik, Dobrovoljc, et al., 2024), with additional metadata fields incorporated to address requirements identified across the various disciplines that make use of such data. The final lists of speaker metadata and recording metadata are provided in Appendices B and C.

3.4 Online interface for data submission

The submission of recordings is facilitated through an online interface, which categorizes users into three distinct roles: participating citizens, administrators, and transcribers. Participating citizens—individuals submitting recordings—are required to register before contributing. For each submission, they must complete metadata forms detailing information about both the speaker and the recorded speech, upload an image of the speaker’s signed consent, and submit the recording. The administrators are responsible for overseeing the data and recordings. They have the ability to download and approve submissions, edit the metadata associated with recordings and speakers, and export summarized spreadsheet data pertaining to registered users, recordings, and speakers. Transcribers are tasked with processing the submitted recordings. They can download audio files and upload transcriptions, which may be provided in both literal and standardized forms as separate entries.

The online interface for data submission is implemented as a web application based on the Flask framework and is hosted under the official domain of the University of Maribor,⁷ ensuring institutional reliability and long-term maintainability. Flask was selected due to its modular architecture, flexibility, and suitability for developing lightweight yet scalable web services. The application integrates a relational database that stores all user-related data, including user profiles, consent records, metadata associated with individual recordings, and system logs. To ensure transparency and traceability, the system implements comprehensive versioning mechanisms: all edits to metadata or user-submitted information are recorded in change log files, allowing complete backtracking and auditability of modifications over time.

Speech recordings and other uploaded files are stored securely and separately from the main application database on the server, which reduces the risk of data corruption and contributes to overall system robustness and maintainability. This architectural separation also facilitates more fine-grained access control and supports compliance with data protection and security requirements, particularly with respect to personal and sensitive data contained in speech recordings. The platform architecture has been designed with scalability in mind, allowing for the future integration of additional annotation layers, processing modules, and corpus-specific workflows. A graphical diagram of the online interface is provided in the appendix A.

Practical use of the portal during its first year of operation has revealed several limitations of the existing administrative interface. Originally designed primarily for internal use, the interface does not sufficiently support efficient data management in scenarios involving broader participation by external users. In particular, the need has emerged for clearer separation of data, a more flexible access management system across different corpora, users, and research projects, as well as a configurable metadata framework capable of accommodating different file types and corpus designs. To address these challenges, a comprehensive upgrade of the administrative interface, underlying data structures, and core functionalities is planned.

The planned upgrade will significantly enhance system stability, accessibility, and reliability, while enabling administrators to flexibly define file types, metadata schemas, and input forms directly through

⁷<https://govorjena-slo-portal.ietk.um.si/>

the web interface, without requiring technical intervention at the code level. This functionality will allow corpus administrators without advanced technical expertise to manage and adapt the platform independently. The upgrade includes the implementation of a full-featured system for managing users, projects or corpora, and uploaded files, as well as mechanisms for assigning access rights based on corpus affiliation and user roles. Extended functionalities will support secure uploading, editing, and storage of recordings, alongside reinforced safeguards for the protection of personal data.

In addition, the upgrade process will encompass systematic testing of the platform, including evaluations with external users, and the preparation of clear and comprehensive user documentation to ensure a user-friendly experience and reliable operation. The resulting infrastructure will establish a solid technical foundation for the sustainable development of spoken language resources, their long-term research and pedagogical use, and increased user involvement in data collection and management processes. In line with principles of openness and reproducibility, the full source code of the platform will be made publicly available upon completion of the project through the CLARIN.SI GitHub repository,⁸ thereby enabling reuse, inspection, and further development by the wider research community.

3.5 Transcription standardization and digital support

The transcription guidelines applied in earlier Slovene speech collection initiatives (Verdonik, Bizjak, et al., 2024; Verdonik, Dobrovoljc, et al., 2024; Verdonik et al., 2013) have been systematically revised, updated, and consolidated on the Govorjena slovenščina platform.⁹ The guidelines are presented in a clearly structured and accessible format, designed both for experienced transcribers and for contributors with limited prior exposure to speech transcription. They are organized into thematically coherent sections covering transcription tools, segmentation principles, literal transcription, standardized transcription, punctuation, transcription-specific conventions, additional annotation marks, and common transcription errors. This modular structure allows users to easily navigate between general principles and more detailed instructions, thereby supporting consistency and transparency in transcription practices.

The platform recommends the use of established transcription tools such as Transcriber (Barras et al., 2001) and EXMARaLDA (Schmidt & Wörner, 2014), both of which support time-aligned transcription and are well suited for spoken corpus compilation. Segmentation into utterances is explicitly advised in order to enhance readability and analytical clarity. Utterances are defined as semantically coherent units, typically corresponding to statements or sentences, rather than being determined solely by prosodic or syntactic criteria. This approach facilitates subsequent linguistic annotation and analysis across different levels. In cases of overlapping speech, transcription is carried out at the word level, and overlapping segments are separated regardless of higher-level utterance boundaries, allowing precise representation of interactional dynamics.

The transcription standard follows a well-established tradition in Slovene spoken language corpora and is based on a dual-mode approach. Literal transcription requires that words be transcribed exactly as pronounced, including reductions, assimilations, and non-standard forms typical of spontaneous speech. In parallel, standardized transcription provides a normalized representation in which each word form is rendered in its standard orthographic form, enabling comparability with written corpora and supporting downstream applications such as lexicographic analysis and language technology development (Verdonik, Bizjak, et al., 2024). Both transcription layers are aligned at the word level, ensuring systematic correspondence between spoken realization and standardized representation.

Punctuation is obligatory and follows conventional written Slovene usage, with sentences beginning with uppercase letters. However, punctuation is applied with sensitivity to spoken discourse structure rather than strict written syntax. The guidelines also define a set of special transcription tags to handle recurrent phenomena in spontaneous speech, including incomprehensible segments, laughter, non-verbal sounds, foreign-language insertions, and hesitation phenomena. In addition, explicit procedures are provided for the anonymization of personal data, such as names, locations, or other identifying information, ensuring compliance with ethical and legal requirements.

⁸<https://github.com/clarinsi>

⁹<https://govorjena-slovenscina.um.si/transkribiranje/>

4 Collected Corpora and Use Cases

The Govorjena slovenščina portal has been in active use since its initial launch in March 2025. During its first year of operation, it has successfully facilitated the collection of several small-scale corpora of Slovene everyday spoken interaction, each serving distinct research and language technology needs. These include (1) the training corpus of spoken Slovene ROG-dialog (Verdonik et al., 2025), (2) the corpus of conversational humour Krohot (Krajnc Ivič et al., 2025), and (3) an addition to the Slovene SloBench benchmark suite for automatic speech recognition evaluation, referred to as SloBenchASR¹⁰.

ROG-dialog was conceived as an extension and enhancement of the existing training corpus of spoken Slovene, ROG-Artur (Verdonik, Ljubešić, et al., 2024), which predominantly comprises public, monologic speech. In contrast, ROG-dialog focuses on informal dialogic interactions, sampling a balanced set of speakers with respect to regional residence, age, and gender to better reflect natural conversational variation. In total, the ROG-dialog corpus comprises 12 recordings involving 25 speakers, resulting in more than 5 hours (50,000 words) of recorded speech. This corpus is openly accessible via the CLARIN.SI repository, supporting reuse in both linguistic research and language technology development.

The Krohot corpus represents a specialized collection of conversational humour. It consists of 10 audio recordings of private, spontaneous conversations involving two or three participants, with a total duration of 3 hours and 52 minutes of natural speech collected primarily between May and September 2025. The recordings capture memories and narratives that elicited spontaneous humorous reactions among interlocutors. Humorous segments within these conversations have been manually annotated using a purpose-built tagging scheme encompassing categories such as vocabulary choice, interpersonal relation, content focus, speaker attitude, and stylistic manner of delivery. The Krohot corpus is provided in high-quality WAV format, with aligned transcriptions available in EXMARaLDA, Transcriber, and plain text formats, and is openly accessible via the CLARIN.SI repository.

Furthermore, Govorjena slovenščina has contributed data to the SloBenchASR component of the SloBENCH evaluation platform, an independent benchmarking framework for Slovenian natural language processing technologies. SloBenchASR provides a dedicated leaderboard for the evaluation of automatic speech recognition systems, using evaluation data that, to the best of current knowledge, are not available in existing Slovene speech databases. The original benchmark dataset consisted of 15 speech recordings with a total duration of 3 hours and 18 minutes, encompassing both public and private speech. This dataset has been expanded through Govorjena slovenščina with the addition of six recordings of private conversational speech, amounting to a total of one hour of additional data. The expanded dataset enables more robust and objective comparison of ASR systems and supports longitudinal tracking of progress in Slovene speech recognition, thereby strengthening the platform’s role in language technology evaluation.

In addition to these purpose-driven collections, more than 25 hours of speech had been contributed to the portal through donated recordings by March 2026. While this reflects an initial stage of community engagement, the project has set a concrete target of collecting at least 70 hours of speech by October 2027. Data collection is conceived as a long-term effort that will continue beyond the project period, with the aim of developing a substantially larger and continuously expanding resource.

In parallel with public participation, the platform is designed to attract researchers and research teams seeking to collect speech data for specific research objectives. By using the portal, such teams can rely on established legal, technical, and organisational frameworks, thereby reducing overhead and ensuring compatibility with existing Slovene speech corpora.

5 Discussion and Conclusion

The persistent underrepresentation of conversational spoken language resources, particularly for everyday human–human interaction, remains a central challenge for both linguistic research and language technology development. As demonstrated throughout this paper, this scarcity cannot be attributed to a single factor, but rather to an interplay of methodological, technical, legal, and social constraints. Tradi-

¹⁰<https://slobench.cjvt.si/leaderboard/view/10>

tional laboratory and fieldwork-based approaches, while indispensable for producing high-quality data, are difficult to scale and are subject to well-known limitations such as the observer’s paradox. At the same time, large-scale online sources provide volume but lack representativeness for private, informal speech. These conditions motivate the exploration of alternative collection paradigms that can balance data quality, diversity, and feasibility.

The analysis of international citizen-science and crowdsourcing initiatives—CorCenCC, Anneta kōnet, and Lahjoita puhetta—illustrates both the potential and the limitations of remote speech collection. These projects demonstrate that large volumes of spoken data can be collected efficiently when technical design, legal governance, and participant motivation are addressed holistically. At the same time, they reveal recurring trade-offs between spontaneity and acoustic control, between short low-effort contributions and longer interactionally rich recordings, and between openness and privacy protection. Crucially, the most successful initiatives treat speech collection platforms as socio-technical systems rather than as purely technical upload tools, foregrounding trust, transparency, and purpose communication alongside usability and quality control.

Within this broader landscape, Govorjena slovenščina occupies a distinct position. Rather than aiming for rapid large-scale data accumulation, the platform is designed as a reusable academic infrastructure that prioritizes legal clarity, methodological consistency, and long-term interoperability of data. Its design choices—such as the absence of a dedicated recording application, the emphasis on longer conversational recordings, contribution limits per speaker, and detailed metadata and transcription standards—reflect a deliberate focus on data depth, contextual richness, and ethical governance. This orientation aligns the platform particularly well with linguistic research that requires extended interactional context, while still supporting language technology applications.

The corpora collected through Govorjena slovenščina—ROG-dialog, Krohot, and the SloBenchASR extension—demonstrate the platform’s capacity to support diverse research goals, ranging from conversational analysis and humor studies to speech technology evaluation. Although the total volume of collected speech remains modest compared to national donation campaigns, these resources validate the platform’s workflows and standards and provide a foundation for future expansion. Early signs of voluntary participation further suggest that, with targeted outreach and sustained engagement strategies, broader community involvement may be achievable over time.

In conclusion, Govorjena slovenščina shows how a standards-driven, legally grounded, and transparent platform can contribute meaningfully to the sustainable development of spoken language resources, even under constrained financial and organizational conditions. Its primary contribution lies not in maximizing data volume, but in lowering structural barriers for researchers and contributors alike by providing a shared infrastructure that integrates legal compliance, technical robustness, and methodological coherence. As the platform continues to evolve—through planned technical upgrades, expanded dissemination efforts, and increased collaboration with research teams—it has the potential to become a central hub for the collection and reuse of spoken Slovene. More broadly, the approach outlined in this paper offers a transferable model for other smaller or under-resourced languages seeking to balance ethical responsibility, research quality, and practical feasibility in the collection of conversational speech data.

6 Acknowledgment

The work presented in this paper was conducted within the research project “Large Language Models for Digital Humanities” (GC-0002) and the Research Programme “Advanced Methods of Interaction in Telecommunication” (P2-0069), both funded by the Slovenian Research and Innovation Agency (ARIS).

A Architectural diagram of the online interface

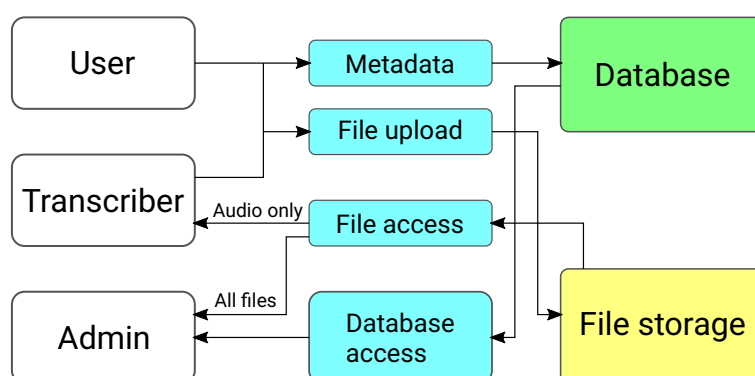


Figure 1: Architectural diagram of the online interface

B Metadata on speakers

- **Transcription ID** – Unique identifier of the transcription linked to the recording.
- **Recording ID (and source ID)** – Identifier of the edited (approved) recording and source recording (in case of renaming).
- **Corpus ID** – Identifier of the corpus to which the recording belongs.
- **Personal ID** – Anonymous identifier assigned to each speaker.
- **Speaker order** – Order in which speakers enter the interaction (e.g. first speaker, second speaker).
- **Gender** – Speaker’s gender.
- **Age** – Speaker’s age at the time of recording, expressed in years, or in months for children up to 6 years old.
- **First language** – Speaker’s primary (native) language.
- **Information on bilingualism** – Whether the speaker has been exposed to and uses more than one language from birth, and in what contexts.
- **Register** – Type of speech variety used (standard, colloquial, dialect).
- **Dialect group** – Broad regional dialect classification.
- **Dialect** – Specific dialect spoken by the speaker.
- **Degree of dialect retention** – Extent to which dialect features are preserved in speech, measured on a scale from 1 to 3.
- **Education** – Speaker’s highest completed level of education.
- **Parent or guardian education (in the case of minors)** – Educational level of the speaker’s parent or guardian who provides consent.
- **Permanent residence** – Current place of residence (settlement, region, country).
- **Childhood residence** – Place where the speaker spent most of their childhood (settlement, region, country).

- **Other residences** – Other places where the speaker has lived for more than a year (settlement, region, country).
- **Atypical speech** – Whether the speech deviates from typical development; if yes, description of the condition.

C Metadata on speech recordings

- **Transcription ID** – Unique identifier of the transcription linked to the recording.
- **Recording ID (and source ID)** – Identifier of the edited (approved) recording and source recording (in case of renaming).
- **Corpus ID** – Identifier of the corpus to which the recording belongs.
- **Language of speech** – Language used in the recording.
- **Manual description of the recorded situation** – Short description of the communicative situation.
- **Date** – Date when the recording was made (month, year).
- **Source** – Origin of the recording (e.g. own recording, recordings from media sources such as national broadcaster, etc.).
- **Text region** – Name of the settlement (city, town, or village) where the recording was made.
- **Genre** – Communicative genre (e.g. conversation, interview, narration).
- **Speech type** – Type of speech situation classified as public formal, public informal, institutional formal, institutional informal, or private.
- **Channel** – Medium of communication, classified according to the target interlocutors or audience (e.g. in-person, telephone, television, radio, internet audio, internet video).
- **Number of speakers** – Total number of participants in the recording.
- **Keywords** – Descriptive tags indicating topics or content.
- **Recording device** – Description of the device used (e.g. smartphone, recorder).
- **Recording location** – Type of physical setting where the recording was made (e.g. living room, student room, office).
- **Original recording format** – File format of the original recording (e.g. WAV, MP3).
- **Manual rating of recording quality** – Subjective assessment of audio/video quality, measured on a scale from 1 to 5.
- **Length of the recording** – Total duration of the recording.
- **Additional field notes** – Any relevant observations about the recording process, participant interaction and behaviour, or notable linguistic and speech features.

References

- Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1), 5–22. [https://doi.org/10.1016/S0167-6393\(00\)00067-4](https://doi.org/10.1016/S0167-6393(00)00067-4)
- Bizjak, A. (2025). Vloga občanske znanosti pri množičnem zbiranju govornih virov v slovenščini. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, 13(1), 58–103. <https://doi.org/10.4312/slo2.0.2025.1.58-103>
- Eskenazi, M., Levow, G.-A., Meng, H., Parent, G., & Suendermann, D. (2013). *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment*. John Wiley & Sons.
- Knight, D., Loizides, F., Neale, S., Anthony, L., & Spasić, I. (2001). Developing computational infrastructure for the CorCenCC corpus: The national corpus of contemporary Welsh. *Language Resources and Evaluation*, 55(3), 789–816. <https://doi.org/10.1007/s10579-020-09501-9>
- Knight, D., Morris, S., Fitzpatrick, T., Rayson, P., Spasić, I., & Thomas, E. M. (2020). The national corpus of contemporary Welsh: Project report — y corpws cenedlaethol cymraeg cyfoes: Adroddiad y prosiect. <https://doi.org/10.48550/arXiv.2010.05542>
- Krajnc Ivič, M., Mihailović, L., Ivič, D., & Verdonik, D. (2025). Corpus of conversational humor Krohot 1.0 [Slovenian language resource repository CLARIN.SI]. <http://hdl.handle.net/11356/2065>
- Labov, W. (1972). *Sociolinguistic patterns (conduct and communication)*. University of Pennsylvania Press.
- Lindén, K., Jauhiainen, T., Lennes, M., Kurimo, M., Rossi, A., Kurki, T., & Pitkänen, O. (2022, October). Donate speech. In D. Fišer & A. Witt (Eds.), *Clarín: The infrastructure for language resources* (pp. 481–510). De Gruyter. <https://doi.org/10.1515/9783110767377-019>
- Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3), 319–344. <https://doi.org/10.1075/ijcl.22.3.02lov>
- Lyding, V., Nicolas, L., & König, A. (2022, June). About the applicability of combining implicit crowdsourcing and language learning for the collection of nlp datasets. In C. Callison-Burch, C. Cieri, J. Fiumara, & M. Liberman (Eds.), *Proceedings of the 2nd workshop on novel incentives in data collection from people: Models, implementations, challenges and results within LREC 2022* (pp. 46–57). European Language Resources Association. <https://aclanthology.org/2022.nidcp-1.8>
- Neale, S., Spasic, I., Needs, J., Watkins, G., Morris, S., Fitzpatrick, T., Marshall, L., & Knight, D. (2017). The CorCenCC crowdsourcing app: A bespoke tool for the user-driven creation of the national corpus of contemporary Welsh. *The 9th International Natural Language Generation conference*, 24–28. <https://eprints.ncl.ac.uk/242273>
- Nielsen, J. (1994, April). Enhancing the explanatory power of usability heuristics. In B. Adelson, S. Dumais, & J. Olson (Eds.), *Proceedings of the sigchi conference on human factors in computing systems* (pp. 152–158). Association for Computing Machinery. <https://doi.org/10.1145/191666.191729>
- Oostdijk, N. (2000, May). The spoken Dutch corpus. Overview and first evaluation. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, & G. Stainhauer (Eds.), *Proceedings of the second international conference on language resources and evaluation (LREC'00)*. European Language Resources Association (ELRA). <https://aclanthology.org/L00-1083/>
- Rutten, M., Minkman, E., & van der Sanden, M. (2017). How to get and keep citizens involved in mobile crowd sensing for water management? A review of key success factors and motivational aspects. *Wiley Interdisciplinary Reviews: Water*, 4(4), e1218. <https://doi.org/10.1002/wat2.1218>
- Schmidt, T., & Wörner, K. (2014). “EXMARaLDA”. In J. Durand, U. Gut, & G. Kristoffersen (Eds.), *Handbook on corpus phonology* (pp. 402–419). Oxford Academic. <https://doi.org/10.1093/oxfordhb/9780199571932.013.030>

- Verdonik, D., Bizjak, A., Žgank, A., Maučec, M. S., Trojar, M., Gros, J. Ž., Bajec, M., Bajec, I. L., & Dobrišek, S. (2024). Strategies for managing time and costs in speech corpus creation: Insights from the Slovenian ARTUR corpus. *Language Resources and Evaluation*, 59, 1899–1924. <https://doi.org/10.1007/s10579-024-09792-2>
- Verdonik, D., Dobrovoljc, K., Erjavec, T., & Ljubešić, N. (2024, May). Gos 2: A new reference corpus of spoken Slovenian. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)* (pp. 7825–7830). ELRA; ICCL. <https://aclanthology.org/2024.lrec-main.691/>
- Verdonik, D., Kosem, I., & Vitez, A. Z. (2013). Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. *Language Resources and Evaluation*, 47(4), 1031–1048. <https://doi.org/10.1007/s10579-013-9216-5>
- Verdonik, D., Ljubešić, N., Rupnik, P., Dobrovoljc, K., & Čibej, J. (2024, September). Izbor in urejanje gradiv za učni korpus govorne slovenščine ROG. In Š. A. Holdt & T. Erjavec (Eds.), *Conference on language technologies and digital humanities (JT-DH-2024)* (pp. 469–484). Inštitut za novejšo zgodovino. <https://doi.org/10.5281/zenodo.13936425>
- Verdonik, D., Rupnik, P., Vidinič, J., & Ljubešić, N. (2025). Corpus of spoken slovenian ROG-dialog 1.0 [Slovenian language resource repository CLARIN.SI]. <http://hdl.handle.net/11356/2073>