The Icelandic Language Biobank: Data Collection through a Clinical Analysis Platform

Iris Nowenstein[1], **Naizeth Núñez Macías**[2], **Gunnar Thor Örnólfsson**[1], **Stefán Ólafsson**[2], **Bryndís Bergþórsdóttir**[1], **Iðunn Kristínardóttir**[1], **Hinrik Hafsteinsson**[1]
[1]University of Iceland, [2]Reykjavík University
{irisen, gunnarthor, brynberg, idunnkristinar, hinhaf}@hi.is,
{naizeth23, stefanola}@ru.is

Abstract

Recent work on clinical applications of language technology shows considerable potential for people with speech and language symptoms and disorders, including for the diagnosis and monitoring of diseases and disorders as well as the development of novel communication aids. This has resulted in a variety of digital health tools becoming accessible, including personalized automatic speech recognition for disordered speech and the monitoring of disease progression in neurodegeneration through language samples. Currently, these tools are almost exclusively accessible to speakers of high-resource languages. A major hurdle for small, lower-resourced language communities in this context is the creation of clinical language corpora. We describe ongoing efforts to build the necessary infrastructure for clinical speech and language data collection in Iceland through the Icelandic Language Biobank, a resource that leverages collaboration with clinicians and robust linguistically-informed data collection against data scarcity.

Keywords: clinical language corpora, speech and language disorders, Icelandic

1. Introduction

Advances in language technology over the last decade have led to a significant body of research exploring clinical applications of automatic speech and language analysis. In the context of speech and language symptoms and disorders, two main types of applications rely on clinical corpora.

The first is diagnosis and monitoring, primarily in the context of neurodegenerative diseases such as Alzheimer's, Frontotemporal Dementia, Parkinson's and ALS (e.g., Cho et al., 2024, Shellikeri et al., 2024, Cao et al., 2025), where language sample analysis (based on e.g. picture descriptions) might yield cost-effective, person-centered and non-invasive endpoints for early screening and treatment efficacy assessments, including in drug trials (Robin et al., 2023). Automatic language sample analysis additionally has a long tradition in the context of developmental language disorders and is particularly valuable for the evaluation of bi- and multilingual children (Ortiz et al., 2024), despite challenges in successful technological transfer to clinicians (Klatte et al., 2022, Liu et al., 2023).

Although we focus on conditions which affect speech and language in the current paper, it is important to note that automatic speech and language analysis has also shown potential for diagnosis and monitoring in other clinical contexts (e.g. Malgaroli et al., 2023, Lombardo et al., 2025)

The second application consists of new technology for alternative and augmentative communication (AAC) aids, such as personalized voice synthesis and speech recognition for disordered speech (e.g. MacDonald et al., 2021, Hasegawa-Johnson et al., 2024, Hyppa-Martin et al., 2024). In both diagnosis/monitoring and AAC applications, a range of digital health tools have become available to users, but only for English or a few other high-resource languages. García et al. (2023) point to the ubiquity of English in the field of speech and language markers of neurodegeneration and call for linguistically diverse research, as well as equitable access to novel clinical instruments. It is fairly straightforward to extend this call to action to the field of communication aids based on language technology.

The current paper describes our attempt to answer the call for Icelandic, a low-to-medium resource language (Daðason and Loftsson, 2024) in a small language community of approximately 400,000 speakers. This is done through the creation of the Icelandic Language Biobank (ILB). In contrast with the most comprehensive data collection efforts in high-resource languages (e.g. Hasegawa-Johnson et al., 2024, Kourtis, 2025), the ILB will contain language samples for both AAC and diagnosis/monitoring in order to maximize data exploitation. While this increases the data management challenges, we argue that it is a necessary counterweight to the data scarcity facing the Icelandic language community, which is exacerbated in a clinical context.

Data will primarily be collected through a web-based semi-automatic linguistic analysis platform, ALDA (Automatic Linguistic Data Analysis), designed for and co-created with speech-language pathologists/therapists (SLPs). The purpose of this

Proceedings of the 6th RaPID@MENTAI.ai Workshop @ LREC 2026, pages 13–23
12 May 2026.

collaborative process is to ensure successful technological transfer to the clinical context, making the use of our speech and language processing pipeline accessible to clinicians through a user-interface which is tailored to their clinical practice. SLPs are experts in the clinical analysis of speech and language and can therefore benefit greatly from the augmentation of their perceptive and/or manual analysis of language samples through the means of language technology (Klatte et al., 2022, Lian et al., 2025). Similarly, their expertise entails the possibility of direct manual corrections in a clinical context, enhancing analysis quality (hence, the platform is semi-automatic). Creating a platform which combines accessible clinician-centered tools and a data sharing infrastructure could therefore create incentives for durable and sustainable clinician-led data collection, another possible counterweight to data scarcity.

In the current paper, we present the process of building the ILB and SLP-oriented platform, discuss challenges and argue for the crucial role of linguistic knowledge in the endeavor of promoting equitable access to healthcare solutions based on language technology, particularly in the context of language-specific manifestations of disorders and diseases.

Finally, we draw on insights based on corpus linguistics (e.g. Gries, 2010, Wolfer and Koplenig, 2025) and argue that the lack of knowledge on clinically relevant linguistic features for low-to-medium resource languages such as Icelandic can in part be compensated for with the collection of larger and more varied language samples, contra the current trend of decreasing sample length in clinical settings to 1-5 minutes (Petti et al., 2023). Recent ideas about leveraging data from wearables and mobile devices (Kourtis et al., 2019), as has in part been done in the context of language acquisition research (Blom et al., 2023), might therefore be considered particularly beneficial for lower-resource languages. This includes the latest developments of analyzing typing behavior, or "smartphone keyboard input patterns to detect early signs of cognitive impairment" (Samsung Newsroom, 2025).

2. Related work

2.1. State of the art in high-resource contexts

Clinical speech and language corpora, particularly in the context of communication disorders, are not only lacking in less-resourced languages. Even for English, considerable efforts and resources are currently being dedicated to clinical speech and language data collection. We describe two such initiatives below, the Speech Accessibility Project and SpeechDx, as well as the web-based application TELL, which extracts speech and language markers of neurodegeneration and is designed both for clinicians and researchers. We consider these three examples as the current state of the art for clinical speech corpora infrastructure in high-resource languages and use them as a references for the Icelandic Language Biobank and our web-based SLP-oriented analysis platform.

The **Speech Accessibility Project** (Hasegawa-Johnson et al., 2024) is a research initiative funded by Amazon, Apple, Google, Meta, Microsoft and nonprofit organizations. Its aim is to improve automatic speech recognition for non-standard (English) speech by collecting more diverse data through crowdsourcing. Currently, the Speech Accessibility Project collects speech data from paid volunteers (from the U.S., Canada and Puerto Rico) with a variety of speech patterns or disorders and compiles them into an anonymized dataset. Participants can join the project through the web page. As of February 2026, the project had collected at least 1500 hours of recorded speech from more than 1000 participants, including people with Parkinson's disease, Down syndrome, Cerebral Palsy, amyotrophic lateral sclerosis (ALS), and people who have had a stroke. The data contains sentences read out loud that correspond to computer commands, sentences read out loud from (sometimes simplified) novels, and spontaneous speech comprising answers to questions about culture or daily life. Companies and researchers can request access to the corpora created in this project. In addition to the audio files, transcripts, original speech prompts and a subset of the corpora annotated by SLPs are also available. The initiative builds on the experience from Google's Project Euphonia (MacDonald et al., 2021, Tobin and Tomanek, 2022) which laid the groundwork for Project Relate, an Android app with personalized speech recognition for non-standard speech, offering both transcription and resynthesis to aid communication. Importantly, deriving speech markers of diseases and disorders is not one of the aims of the Speech Accessibility Project.

Speech-based biomarkers are on the other hand the main focus of **SpeechDx** (Kourtis, 2025), a project in which the goal is to create a longitudinal dataset (spanning three years) for the diagnosis of Alzheimer's disease and related dementias through speech samples. The goal is to cover English, Spanish and Catalan and to link comprehensive clinical information to the participants' speech data. The dataset includes samples obtained through different elicitation tasks, such as picture descriptions, open-ended questions, story recall and storytelling. The data will be hosted by the Alzheimer's Disease Data Initiative and will be made available to researchers approved by a committee in a pro-

tected, controlled environment.

The **TELL** application (García et al., 2024b) provides "robust speech biomarkers for clinical and research purpose" and has mostly been deployed in the context of neurodegeneration. The first deployment was available for English, Spanish, French and Portuguese. Although its latest version also enables data collection for German, Italian, Quechua, Kiswahili and Tagalog (García et al., 2024a), language-specific features (such as POS tags, semantic granularity etc.) are only (automatically) extracted for the higher-resource languages. The platform is designed to be used by clinicians as well as researchers, but SLPs are not the main target group. This is comparable to Open Brain AI (Themistocleous, 2024), a relatively new computational platform which currently provides tools for automatic language sample analysis in 15 languages and is also designed for both clinicians and researchers.

To the best of our knowledge, no comprehensive automatic speech and language analysis platform is designed for the needs of SLPs both in the fields of developmental and acquired communication disorders. One of the key challenges is the technical skills needed to implement available tools in clinical practice. For instance, the Batchalign pipeline was developed for the automatic transcription and analysis of clinical samples in the CHAT (Codes for the Human Analysis of Talk) format using the software program CLAN (Computerized Language Analysis, Liu et al., 2023). One of the main goals of Batchalign was to reduce the time needed to transcribe raw audio files by enabling clinicians to generate an automatic transcription, which only had to be manually corrected. However, another study found that SLPs did not experience a reduction in the time needed to perform language sample analysis despite receiving tailored training (Klatte et al., 2022). In fact, the participants reported a lack of knowledge and skills as a barrier to using tools such as Batchalign. These results highlight the need for user-friendly software and the incorporation of SLPs in the tool design process. Additionally, we believe a crucial component, particularly for lower-resource languages with fewer tools and higher error rates, is to enable clinicians such as SLPs to correct the automatic analysis.

2.2. The Icelandic landscape

Not unexpectedly, Icelandic implementations within state-of-the-art tools and datasets for digital health are limited. For example, there are no direct analogs to the aforementioned TELL, Open Brain AI and CHAT in Iceland. This is in part due to the lack of datasets, research and development in the domain of clinical language technology for Icelandic.

As is the case in many small(er) language communities, collecting clinical linguistic data for Icelandic presents a number of challenges. This is not only due to the limited number of speakers, but also to the lack of infrastructure to safely collect, store, and share language samples, as well as the lack of focused research on clinical populations speaking those languages. In this sense, there is a real risk that Icelandic will fall further behind with regards to the development and accessibility of automatic speech and language analysis tools and datasets for digital health.

As briefly mentioned, an inherent difficulty for small language communities like Icelandic is the overall low number of individuals diagnosed with the targeted diseases and disorders. For example, there are approximately 20-30 people with ALS in Iceland at any given time (MND Iceland). These individuals face the same difficulties as people with ALS in larger communities, but the development of novel, language-specific solutions is limited by the fact that data collection can only be obtained through a very low number of individuals. We believe this entails that any novel data for Icelandic needs to be utilized to the fullest.

Fortunately, there are already projects that have gathered Icelandic language samples for uses in clinical language technology development. Specifically, these projects gathered language samples from people with Mild Cognitive Impairment and mild dementia due to Alzheimer's disease (Curcic et al., 2022, Callegari et al., 2023 and Nowenstein et al., 2024). However, there are no accessible corpora with language samples from other clinical groups where automatic language sample analysis has proven useful, e.g., people with Parkinson's, ALS, Frontotemporal Dementia and aphasia and/or motor speech disorders following a stroke, as well as language samples from children with developmental language disorders.

The first results derived from the two Alzheimer's disease datasets described above highlight the importance of taking into account the characteristics of individual languages when generalizing previous results and extending the use of clinical language technology to new contexts. For example, Callegari et al. (2024) show that the frequency of various features will vary greatly across discourse contexts in different types of language samples, both within features commonly extracted in previous research (e.g. verb rate) and features more specific to the Icelandic language (e.g. subjunctive rate).

Another recent project comparing speech and language markers of neurodegeneration across English, Korean and Icelandic found that language-specific features, such as the use of case marking in Icelandic, can differentiate between clinical groups and healthy controls (Nowenstein et al.,

2025). Similar results have been found in the field of Developmental Language Disorders (DLD), where research on Icelandic indicates that the influential view (Leonard, 2014) of morphological errors as a hallmark of DLD does not necessarily hold for languages with richer morphological systems (Thordardottir, 2016).

3. Building the Icelandic Language Biobank

The Icelandic Language Biobank (ILB) is a three year initiative funded by The Strategic Research and Development Programme for Language Technology within the Icelandic Research Fund. Its preparation was also funded by the Language Technology Programme for Icelandic. The ILB is an attempt to answer the call for increased linguistic diversity in clinical language technology (García et al., 2023) and could serve as a proof of concept for other small language communities.

Our goal is to collect clinical language samples from speakers of Icelandic (children and adults, mono- and multilingual) in order to improve access to healthcare solutions based on language technology. We focus on the manifestations of developmental language disorders and neurodegeneration in Icelandic and build the infrastructure for the collection and preservation of clinical speech and language corpora for Icelandic in a broad sense. A further goal of the ILB infrastructure is to allow for the collection, annotation, and analysis of samples in other languages.

We collaborate with the leadership of the Speech Accessibility project (Hasegawa-Johnson et al., 2024) and the DELAD initiative within CLARIN. DELAD facilitates the sharing of corpora of speech of individuals with communication disorders among researchers in a GDPR compliant way and at secure repositories in the CLARIN infrastructure (Lee et al., 2024).

The design of the ILB is motivated by the small size and lack of resources of the Icelandic language community. We try to combine the approaches of initiatives such as the Speech Accessibility Project, SpeechDx and TELL into one centralized solution to maximize data exploitation. This means that the same speaker can provide data for improved communication aids (e.g. ASR for disordered speech) and diagnosis/monitoring (e.g. digital biomarkers of neurodegenration) through a clinician (SLP) who has access to automatic speech and language data analysis with our data collection and analysis platform. Additionally, SLPs' expertise in speech and language means they are particularly powerful collaborators, including in the context of transcription and annotation correction.

One of the key motivations behind the ILB is the current lack of knowledge when it comes to the direct clinical application of language technology, particularly when it comes to the interpretability of the information retrieved through automatic speech and language analysis. Even though automatic analysis tools may reduce the workload of SLPs (Liu et al., 2023), it is not always clear how SLPs can interpret these new measures within evidence-based practice (Lindsay et al., 2021, Yeung et al., 2021).

Another area where knowledge is lacking is the transfer of research findings from English to other languages, as it is known that the manifestations of language disorders depend to some extent on the specific characteristics of different languages. For instance, research has found that there is an increase in the rate of pronouns in English in Alzheimer's disease (see Petti et al., 2020, Robin et al., 2021 and Cho et al., 2022), while the reverse pattern was found for pro-drop languages such as Bengali (Bose et al., 2021). An overemphasis on English can therefore produce a biased view of what characterizes language symptoms and disorders in general (Thordardottir, 2016, García et al., 2023), which underlines the need for typologically diverse clinical linguistic data and the necessary infrastructure to collect it. This is relevant, among other things, for improved diagnosis of developmental disorders in multilingual children.

3.1. Infrastructure and data management

Data management is one of the key challenges when building infrastructure for the collection and preservation of clinical speech and language data. The ILB will draw from the various data collection initiatives described in section 2.1, complying with e.g. the GDPR and taking into consideration other developments at the EU level, such as the AI and Data Acts as well as the European Health Data Space Regulation. The project also needs to comply with Icelandic legislature on biobanks and clinical datasets and request approval from The National Bioethics Committee of Iceland. The project is hosted at the University of Iceland and has benefitted from consultations with the university's IT departments as well as legal counsel specialized in personal data protection.

As is detailed in García et al. 2024a, the TELL application is deployed on Amazon Web Services (AWS) with data handling through Amazon's RDS PostgreSQL database, audio files stored in AWS S3 and patient health information encrypted on AWS Key Management Service. Our current data management plan is similar in structure but is currently hosted completely locally at the University of Iceland given Icelandic legislature which dictates that clinical data should in general be exclusively processed and stored domestically. This will poten-

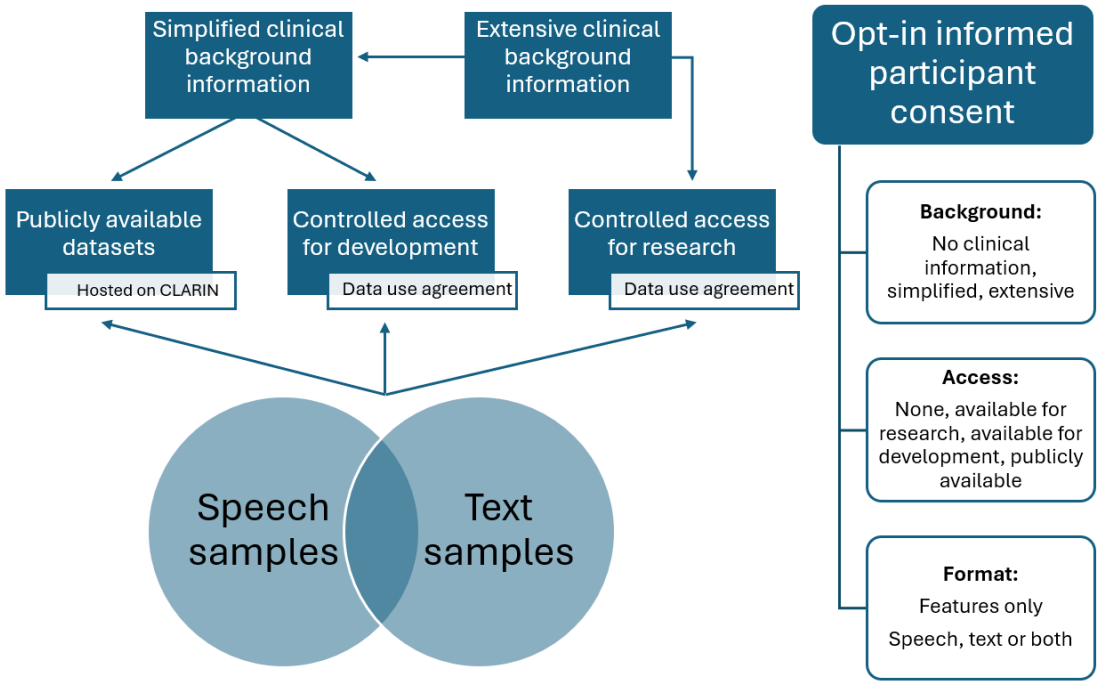

Figure 1: Icelandic Language Biobank consent and data sharing infrastructure.

tially cause scalability challenges with increased data collection efforts, a complication we are currently addressing.

When data has been collected through the web platform, we transfer it to a Nextcloud server hosted within the IREI cluster at the University of Iceland through encrypted data transfer. At the University of Iceland, the data is reviewed and personally identifying information (PII) is manually removed. Participants' identification numbers are stored separately from the deidentified, pseudonymized data and not shared unless a new approval by The National Bioethics Committee has been granted. As shown in Figure 1, the ILB will adopt a layered approach with opt-in (and opt-out) informed participant consent for the deidentified data, meaning that the project will share data in a variety of forms.

This means that participants can opt into and out of different levels of data sharing for different groups, allowing for example developers to have controlled access to simplified or no clinical information through a data use agreement while researchers can access more extensive clinical information, also under a data use agreement. Furthermore, participants will have control over what information from their sample is shared, ranging from audio recordings to feature values only. This entails providing participants with comprehensive information, for example regarding their voice potentially being identifiable to their acquaintances.

Since the ILB aims to collect data from a number of vulnerable groups and/or protected entities, precautions are in order to ensure that the consent can truly be considered informed. In some cases, this will be in the form of parental or guardian informed consent as well as participant assent with a simplified information form. To ensure appropriate informed consent, the ILB builds on the DELAD resources presented in Lee et al. (2024) as well as the PEEC (Protected Entities Ethics Checklist) for collecting speech data from vulnerable clinical populations (Choi et al., forthcoming). For instance, according to the PEEC framework, children's consent should be tailored to their developmental stage. Children above 7 may provide written assent, while the affirmative willingness of younger participants may be provided after a simplified, verbal explanation (Choi et al., forthcoming). At any given point, participants may withdraw their consent and demand the deletion of their data.

Data sharing will be handled under restricted licenses within the DELAD-initiative through CLARIN, as DELAD is linked to CLARIN's Knowledge Centre for Atypical Communication Expertise (ACE) for making corpora of speech of individuals with communication disorders (CSD) available through The Language Archive (TLA) at the Max Planck Institute in Nijmegen (a CLARIN Data Centre) and CMU's Talkbank (Clinical Banks). We therefore provide an infrastructure which makes it possible to ensure the data of the ILB will be as accessible as possible while guaranteeing participants' data privacy according to their wishes.

3.2. ALDA: web-based semi-automatic linguistic analysis platform

A key driver of sustainable and longitudinal collection of clinical data within the ILB is collaboration with SLPs and other clinicians. For such collaboration to be gainful for all parties, we strive to provide utility to the clinicians in exchange for their contribution to data collection. Therefore, the development team for the ALDA platform includes two SLPs, one working with pediatric populations and the other specializing in neurodegeneration.

ALDA (Automatic Linguistic Data Analysis) is a web-based semi-automatic linguistic analysis platform that allows clinicians specialized in speech and language disorders to perform Language Technology-aided analysis of clinical language samples through a user-friendly interface. The platform also allows the clinicians to manage their clients' language samples in a centralized storage space and track indicators of diseases and disorders over time. The platform is currently being developed with funding from the Language Technology Programme for Icelandic.

As illustrated in Figure 2, the SLP records a speech sample using ALDA, for example a picture description or story recall, or uploads a previous recording. The sample is then transcribed and diarized using automatic speech recognition and corrected manually as necessary. This step of language sample analysis has so far been performed fully manually by Icelandic SLPs. ALDA also supports tasks for children and adults with feature bundles for different types of speech and language disorders.

ALDA utilizes various speech and language processing tools to perform automated analyses of the speech sample, e.g. for ASR and speaker diarization, POS-tagging and parsing. The current version of the platform integrates WhisperX (Bain et al., 2023) for VAD (Voice Activity Detection), ASR, forced alignment and speaker diarization. For optimal results, versions of Whisper and Wav2Vec2 which have been fine-tuned on Icelandic speech corpora are used (Radford et al., 2023, Baevski et al., 2020, Mena et al., 2024). A PoS tagger using a fine-grained morphological tagset (Jónsson et al., 2021) is used to extract morphosyntactic information from the transcribed speech. We stress that the pipeline is under development and needs further testing in terms of e.g. preprocessing steps with diverse data sources as well as error rates for children's voices and disordered speech. Our text processing pipeline will also keep evolving. Although preliminary findings suggest acceptable POS-tagging accuracy in clinical conversational language samples, a lot of challenges remain for syntactic parsing (not unexpectedly, see Agmon et al., 2026 for English). Finally, the current pipeline only supports the analysis of Icelandic language samples.

After data collection, recordings and analysis results can then be saved in the SLPs' secure storage space within the platform. ALDA also has a built-in data collection functionality that enables SLPs using the platform to invite clients to receive information about the ILB project. When a client is interested in participating and informed consent has been obtained through an electronic signature on the web page connected to the Icelandic Language Biobank, the existing participant's language samples can be transferred to the ILB. This data collection sets the ILB apart methodologically, as the data collection tool itself is an aid in the current workflow of SLPs who already record language samples within their clinical practice. With ALDA, SLPs have a tool for immediate use in practice, not only after the data have been processed for further research and technology transfer.

In order to involve further SLPs into the development process, two focus groups with Icelandic SLPs have been conducted, confirming the group's interest in using language technology to facilitate language sample analysis. The participants also expressed enthusiasm for the use of language technology for communication aids and were positive towards collaboration with researchers, but emphasized the importance of receiving detailed instructions and a clear and accessible protocol. Interestingly, SLPs did not show enthusiasm for language sample analysis in languages they do not speak and cited the importance of being able to verify the analysis to interpret the results.

In addition to the focus groups, 35 practicing SLPs participated in a survey we conducted about language sample analysis. 91.4% of the participants considered language samples as a useful tool, the rest considered it useful in certain cases. Additionally, 91% of our participants said they would use language samples more frequently if processing them took less time and they had access to better tools for it. Currently, user testing for the platform is ongoing. Until the end of 2026, we aim to identify existing barriers, technical difficulties and SLPs' concerns in order to iteratively improve the platform before starting data collection. Although the current pipeline only supports Icelandic, the platform can of course be used to record and store samples in other languages.

3.3. Data collection within the ILB

In the second year of the Icelandic Language Biobank project (January 2027), its infrastructure will be ready and an initial data collection and analysis phase will begin. Within the project, data collection will be conducted through clinician-led participant recruitment as well as crowdsourcing efforts.

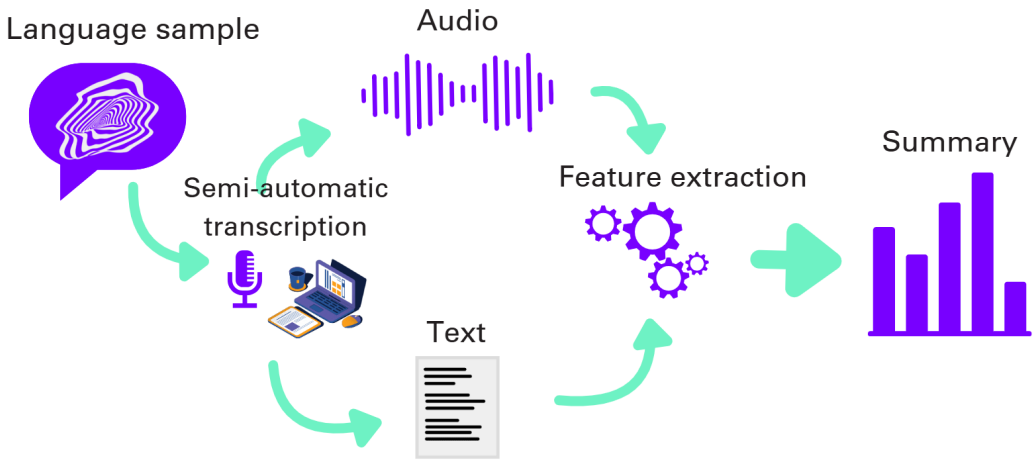

Figure 2: High-level diagram of the flow of language samples through the analysis pipeline. Features are extracted from both the diarized, text-aligned recording and the text transcript.

Both the target dataset size and characteristics of participants will be established during the first three months of data collection based on funding availability and clinician adoption of the web platform, but the primary goal of the project is to build the necessary infrastructure for continuous data collection.

The crowdsourcing will target clinical groups (as has been done in the Speech Accessibility Project) but mainly people without communication disorders. This is because having robust control group data is a crucial step for the use of automatic speech and language analysis in a clinical setting, where situating individual patients within well-established norms is an important component of diagnosis. Indeed, 88% of participants in the aforementioned SLP survey said they would use language samples more if they had better norms as a comparison.

The clinician-led recruitment of participants from clinical groups will be based on the use of the ALDA platform, with the ILB project providing training for SLPs in the form of in-person courses and online instructional content. The first phase of data collection within the ILB project will be centered around developmental language disorders in mono- and multilingual children as well as speech and language disorders in neurodegeneration, using classic language sampling methods such as picture descriptions and story recall.

Although the initial data collection is focused on clinician-led recruitment and crowdsourcing, the goal of the ILB project still is to establish a durable infrastructure for any type of clinical speech and language data, including new types of language samples collected through e.g. wearables and mobile devices. Considering common knowledge about the positive effects of increasing corpus size to obtain more representative samples of language use (e.g. Gries, 2010), we believe that efforts should be made to increase clinical language sample length without increasing clinician burden. Currently, the direction in the literature is to decrease sample length (Petti et al., 2023), but larger clinical language samples from individuals might be an important way of countering the data scarcity inherent to smaller language communities. We also believe there might be benefits to expanding clinical language sampling to analyses of participants' written language output and the ILB will therefore also accommodate the storage of written language samples. Scaling from relatively homogeneous, short spoken language samples to more diverse and extensive types of data will present problems we have not solved yet but aim to address, both in terms of storing and processing/analyzing the data. Nevertheless, we believe academic research in small language communities should strive to accommodate as much data diversity as possible.

4. Conclusion

Recent applications of language technology show that children and adults with communication disorders and the clinicians who treat them stand to benefit considerably from successful technological transfer of speech and language processing tools. A necessary step in that process is the collection of speech and language samples from people with communication disorders. In the context of less-resourced languages, particularly in small language communities where data scarcity will be problematic, we suggest building data infrastructure which will make it possible to collect data both for (1) diagnosis/monitoring of diseases and disorders (including clinical information about the participants) and (2) communication aids, including better

speech recognition for disordered speech.

We presented our approach for this kind of infrastructure in Iceland through the creation of the Icelandic Language Biobank, a project which also includes comprehensive collaboration with clinicians by providing them with tools for data analysis on a platform which additionally serves as a data collection point. We believe that this combination of a comprehensive one-stop approach to corpora of speech of individuals with communication disorders and bilateral collaboration with clinicians will provide some of the necessary counterweight to data scarcity in small less-resourced language communities.

Another way to approach the problem is through longer clinical language samples, possibly leveraging wearables and mobile devices and going beyond speech samples to include individuals' written outputs as well. For this kind of research on large clinical language samples, it is even more important to build data infrastructure where data security and personal privacy are guaranteed.

5. Acknowledgements

The projects presented in the current paper were funded by The Strategic Research and Development Programme for Language Technology within the Icelandic Research Fund and the Language Technology Programme for Icelandic. We would also like to thank the anonymous reviewers for their valuable feedback and comments.

6. Ethical considerations

Building the Icelandic Language Biobank relies on an in-depth mapping of ethical consideration. This is why we base our work on the Protected Entities Ethics Checklist (PEEC, Choi et al., forthcoming), a comprehensive framework specifically designed for researchers collecting speech and language data from clinically vulnerable populations, including children, elderly adults with cognitive changes, individuals with communication disorders, and marginalized communities.

Finally, although the motivation for the Icelandic Language Biobank is centered around stakeholder needs (ensuring accessibility to clinical language technology regardless of the language people speak), it still is crucial to consult with stakeholders, particularly in vulnerable clinical populations, to ensure their perspectives and interests are embedded into the project they contribute to.

7. Bibliographical References

Galit Agmon, Sunghye Cho, Sharon Ash, Katheryn A. Q. Cousins, Kaj Blennow, Henrik Zetterberg, Leslie M. Shaw, Sameer Pradhan, Yoon Duk Kim, Mark Y. Liberman, David J. Irwin, and Naomi Nevler. 2026. Automatic quantification of syntactic complexity in natural spontaneous speech of people with primary progressive aphasia. *Aphasiology*, 40(3):561–582.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. In *Interspeech 2023*, pages 4489–4493.

Elma Blom, Paula Fikkert, Annette Scheper, Merel van Witteloostuijn, and Petra van Alphen. 2023. The Language Environment at Home of Children With (a Suspicion of) a Developmental Language Disorder and Relations With Standardized Language Measures. *Journal of Speech, Language, and Hearing research*, 66(8):2821–2830.

Arpita Bose, Niladri S. Dash, Samrah Ahmed, Manaswita Dutta, Aparna Dutt, Ranita Nandi, Yesi Cheng, and Tina M. D. Mello. 2021. Connected Speech Characteristics of Bengali Speakers With Alzheimer's Disease: Evidence for Language-Specific Diagnostic Markers. *Frontiers Aging Neuroscience*, 13:707628.

Elena Callegari, Iris Edda Nowenstein, Ingunn Jóhanna Kristjánsdóttir, and Anton Karl Ingason. 2024. Automatic Extraction of Language-Specific Biomarkers of Healthy Aging in Icelandic. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1915–1924, Torino, Italia. ELRA and ICCL.

Elena Callegari, Agnes Sólmundsdóttir, and Anton Karl Ingason. 2023. The ACoDe Project: Creating a Dementia Corpus for Icelandic. In *Proceedings of CLARIN Annual Conference 2023*, pages 100–105.

Fangyuan Cao, Adam P. Vogel, Puya Gharahkhani, and Miguel E. Renteria. 2025. Speech and language biomarkers for parkinson's disease prediction, early diagnosis and progression. *npj Parkinson's Disease*, 11(1):57.

Sunghye Cho, Katheryn Alexandra Quilico Cousins, Sanjana Shellikeri, Sharon Ash, David John Irwin, Mark Yoffe Liberman, Murray Grossman, and Naomi Nevler. 2022. Lexical and Acoustic Speech Features Relating to Alzheimer Disease Pathology. *Neurology*, 99(4):e313–e322.

Sunghye Cho, Christopher A. Olm, Sharon Ash, Sanjana Shellikeri, Galit Agmon, Katheryn A. Q. Cousins, David J. Irwin, Murray Grossman, Mark Liberman, and Naomi Nevler. 2024. Automatic classification of AD pathology in FTD phenotypes using natural speech. *Alzheimer's & Dementia*, 20(5):3416–3428.

Anna Seo Gyeong Choi, Sunghye Cho, and Iris Nowenstein. PEEC: The Protected Entities Ethics Checklist for Collecting Speech Data from Vulnerable Clinical Populations. Accepted manuscript, *Journal of Speech, Language, and Hearing Research*.

Claire Cordella, Manuel J. Marte, Hantian Liu, and Swathi Kiran. 2025. An Introduction to Machine Learning for Speech-Language Pathologists: Concepts, Terminology, and Emerging Applications. *Perspectives of the ASHA Special Interest Groups*, 10(2):432–450.

Jelena Curcic, Vanessa Vallejo, Jennifer Sorinas, Oleksandr Sverdlov, Jens Praestgaard, Mateusz Piksa, Mark Deurinck, Gul Erdemli, Maximilian Bügler, Ioannis Tarnanas, Nick Taptiklis, Francesca Cormack, Rebekka Anker, Fabien Massé, William Souillard-Mandar, Nathan Intrator, Lior Molcho, Erica Madero, Nicholas Bott, Mieko Chambers, Josef Tamory, Matias Shulz, Gerardo Fernandez, William Simpson, Jessica Robin, Jón G. Snædal, Jang-Ho Cha, and Kristin Hannesdottir. 2022. Description of the Method for Evaluating Digital Endpoints in Alzheimer Disease Study: Protocol for an Exploratory, Cross-sectional Study. *JMIR Research Protocols*, 11(8):e35442.

Jón Daðason and Hrafn Loftsson. 2024. Text filtering classifiers for medium-resource languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15789–15801, Torino, Italia. ELRA and ICCL.

Adolfo M. García, Jessica de Leon, Boon Lead Tee, Damián E. Blasi, and Maria Luisa Gorno-Tempini. 2023. Speech and language markers of neurodegeneration: a call for global equity. *Brain*, 146(12):4870–4879.

Adolfo M. García, Franco J. Ferrante, Gonzalo Pérez, Joaquín Ponferrada, Alejandro Sosa Welford, Nicolás Pelella, Matías Caccia, Laouen Mayal Louan Belloli, Cecilia Calcaterra, Catalina González Santibáñez, Raúl Echegoyen, Mariano Javier Cerrutti, Fernando Johann, Eugenia Hesse, and Facundo Carrillo. 2024a. Toolkit to Examine Lifelike Language v.2.0: Optimizing Speech Biomarkers of Neurodegeneration. *Dementia and Geriatric Cognitive Disorders*, 54(2):96–108.

Adolfo M. García, Fernando Johann, Raúl Echegoyen, Cecilia Calcaterra, Pablo Riera, Laouen Belloli, and Facundo Carrillo. 2024b. Toolkit to Examine Lifelike Language (TELL): An app to capture speech and language markers of neurodegeneration. *Behavior Research Methods*, 56(4):2886–2900.

Stefan Th. Gries. 2010. Useful statistics for corpus linguistics. In Aquilino Sánchez and Moisés Almela, editors, *A mosaic of corpus linguistics: selected approaches*, pages 269–291. Peter Lang, Frankfurt.

Mark Hasegawa-Johnson, Xiuwen Zheng, Heejin Kim, Clarion Mendes, Meg Dickinson, Erik Hege, Chris Zwilling, Marie Moore Channell, Laura Mattie, Heather Hodges, Lorraine Ramig, Mary Bellard, Mike Shebanek, Leda Sarı, Kaustubh Kalgaonkar, David Frerichs, Jeffrey P. Bigham, Leah Findlater, Colin Lea, Sarah Herrlinger, Peter Korn, Shadi Abou-Zahra, Rus Heywood, Katrin Tomanek, and Bob MacDonald. 2024. Community-Supported Shared Infrastructure in Support of Speech Accessibility. *Journal of Speech, Language, and Hearing Research*, 67(11):4162–4175.

Jolene Hyppa-Martin, Jason Lilley, Mo Chen, Jaclyn Friese, Corinne Schmidt, and H Timothy Bunnell. 2024. A large-scale comparison of two voice synthesis techniques on intelligibility, naturalness, preferences, and attitudes toward voices banked by individuals with amyotrophic lateral sclerosis. *Augmentative and Alternative Communication*, 40(1):31–45.

Inge S. Klatte, Vera Van Heugten, Rob Zwitserlood, and Ellen Gerrits. 2022. Language Sample Analysis in Clinical Practice: Speech-Language Pathologists' Barriers, Facilitators, and Needs. *Language, Speech, and Hearing Services in Schools*, 53(1):1–16.

Lampros C. Kourtis. 2025. Speechdx: A gold-standard speech-and-language dataset for prognostic ad biomarker development. *Alzheimer's & Dementia*, 21:e104638.

Lampros C. Kourtis, Oliver B. Regele, Justin M. Wright, and Graham B. Jones. 2019. Digital

biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. *npj Digital Medicine*, 2(1):9.

Alice Lee, Nicola Bessell, Henk Van Den Heuvel, Katarzyna Klessa, and Satu Saalasti. 2024. The DELAD initiative for sharing language resources on speech disorders. *Language Resources and Evaluation*, 58(3):865–879.

Laurence B. Leonard. 2014. Children with specific language impairment and their contribution to the study of language development. *Journal of Child Language*, 41(S1):38–47.

Jiachen Lian, Xuanru Zhou, Chenxu Guo, Zongli Ye, Zoe Ezzes, Jet M.J. Vonk, Brittany Morin, David Baquirin, Zachary Miller, Maria Luisa Gorno-Tempini, and Gopala Krishna Anumanchipalli. 2025. Automatic Detection of Articulatory-Based Disfluencies in Primary Progressive Aphasia. *IEEE Journal of Selected Topics in Signal Processing*, 19(5):810–826.

Hali Lindsay, Johannes Tröger, and Alexandra König. 2021. Language Impairment in Alzheimer's Disease-Robust and Explainable Evidence for AD-Related Deterioration of Spontaneous Speech Through Multilingual Machine Learning. *Frontiers in Aging Neuroscience*, 13:642033.

Houjun Liu, Brian MacWhinney, Davida Fromm, and Alyssa Lanzi. 2023. Automation of Language Sample Analysis. *Journal of Speech, Language, and Hearing Research*, 66(7):2421–2433.

Clara Lombardo, Giulia Esposito, Silvia Carbone, Salvatore Serrano, and Carmela Mento. 2025. Speech analysis and speech emotion recognition in mental disease: a scoping review. *Frontiers in Psychology*, 16.

Robert L. MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn A. Ladewig, Jimmy Tobin, Michael P. Brenner, Philip C. Nelson, Jordan R. Green, and Katrin Tomanek. 2021. Disordered Speech Data Collection: Lessons Learned at 1 Million Utterances from Project Euphonia. In *Interspeech 2021*, pages 4833–4837. ISCA.

Matteo Malgaroli, Thomas D Hull, James M Zech, and Tim Althoff. 2023. Natural language processing for mental health interventions: a systematic review and research framework. *Translational Psychiatry*, 13(1):309.

Carlos Mena, Þorsteinn Daði Gunnarsson, and Jon Gudnason. 2024. SamróMur MilljóN: An ASR corpus of one million verified read prompts in Icelandic. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14305–14312, Torino, Italia. ELRA and ICCL.

MND Iceland. Hvað er MND? [What is ALS?]. Accessed October 24, 2025.

Iris Nowenstein, Min Seok Baek, Bryndís Bergþórsdóttir, Daria Birju, Elena Callegari, Hinrik Hafsteinsson, Anton Karl Ingason, María K. Jónsdóttir, Ashley Keaton, Sungoo Kim, Seohee Kim, Louis Kwak, Judith Neugroschl, Caitlin Richter, Mary Sano, Truda Silberstein, Jón Snædal, Laila Soleimani, Gunnar Thor Örnólfsson, Carolyn Zhu, and Sunghye Cho. 2025. Speech and language markers of cognitive decline and neurodegeneration: Generalizability across languages. In *Society for the Neurobiology of Language 17th Annual Meeting*, Gallaudet University.

Iris Nowenstein, Marija Stanojevic, Gunnar Örnólfsson, María Kristín Jónsdóttir, Bill Simpson, Jennifer Sorinas Nerin, Bryndís Bergþórsdóttir, Kristín Hannesdóttir, Jekaterina Novikova, and Jelena Curcic. 2024. Speech and Language Biomarkers of Neurodegenerative Conditions: Developing Cross-Linguistically Valid Tools for Automatic Analysis. In *Proceedings of the Fifth Workshop on Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments @LREC-COLING 2024*, pages 26–33, Torino, Italia. ELRA and ICCL.

José A. Ortiz, Jessica M. Nolasco, Yi Ting Huang, and Jason C. Chow. 2024. The Use of Language Sample Analysis to Differentiate Developmental Language Disorder From Typical Language in Bilingual Children: A Systematic Review and Meta-Analysis. *Journal of Speech, Language, and Hearing Research*, 67(10):3803–3825.

Ulla Petti, Simon Baker, and Anna Korhonen. 2020. A systematic literature review of automatic Alzheimer's disease detection from speech and language. *Journal of the American Medical Informatics Association*, 27(11):1784–1797.

Ulla Petti, Simon Baker, Anna Korhonen, and Jessica Robin. 2023. How Much Speech Data Is Needed for Tracking Language Change in Alzheimer's Disease? A Comparison of Random Length, 5-Min, and 1-Min Spontaneous Speech Samples. *Digital Biomarkers*, 7(1):157–166.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever.

2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Jessica Robin, Mengdan Xu, Aparna Balagopalan, Jekaterina Novikova, Laura Kahn, Abdi Oday, Mohsen Hejrati, Somaye Hashemifar, Mohammadreza Negahdar, William Simpson, and Edmond Teng. 2023. Automated detection of progressive speech changes in early alzheimer's disease. *Alzheimer's & Dementia*, 15(2):e12445.

Jessica Robin, Mengdan Xu, Liam D. Kaufman, and William Simpson. 2021. Using Digital Speech Assessments to Detect Early Signs of Cognitive Impairment. *Frontiers in Digital Health*, 3:749758.

Samsung Newsroom. 2025. [World Alzheimer's Day] Samsung Research Advances Early Detection of Alzheimer's With Everyday Digital Data.

Sanjana Shellikeri, Sunghye Cho, Sharon Ash, Carmen Gonzalez-Recober, Katheryn A Q Cousins, Corey T McMillan, Lauren Elman, Colin Quinn, Defne A Amado, Michael Baer, David J Irwin, Lauren Massimo, Mark Y Liberman, and Naomi Nevler. 2024. Digital speech markers of cognitive impairment in ALS-FTD spectrum disorders. *Alzheimer's & Dementia*, 20(S2):e089943.

Charalambos Themistocleous. 2024. Open Brain AI and language assessment. *Frontiers in Human Neuroscience*, 18:1421435.

Elin Thordardottir. 2016. Grammatical morphology is not a sensitive marker of language impairment in Icelandic in children aged 4–14 years. *Journal of Communication Disorders*, 62:82–100.

Jimmy Tobin and Katrin Tomanek. 2022. Personalized automatic speech recognition trained on small disordered speech datasets. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6637–6641.

Sascha Wolfer and Alexander Koplenig. 2025. Does corpus size influence normalised frequencies? *Corpus Linguistics and Linguistic Theory*.

Anthony Yeung, Andrea Iaboni, Elizabeth Rochon, Monica Lavoie, Calvin Santiago, Maria Yancheva, Jekaterina Novikova, Mengdan Xu, Jessica Robin, Liam D. Kaufman, and Fariya Mostafa. 2021. Correlating natural language processing and automated speech analysis with clinician assessment to quantify speech-language changes in mild cognitive impairment and Alzheimer's dementia. *Alzheimer's Research & Therapy*, 13(1):109.

8. Language Resource References

Jónsson, Haukur Páll and Loftson, Hrafn and Steingrímsson, Steinþór. 2021. *ABLTagger (PoS) - 3.0.0*. Reykjavík University. PID http://hdl.handle.net/20.500.12537/115. CLARIN-IS.