A shoal of voices: Parallel read speech from professional Swedish narrators

Christina Tånnander, Jim O'Regan, Jens Edlund
Swedish Agency for Accessible Media (MTM), KTH Royal Institute of Technology
Stockholm, Sweden
christina.tannander@mtm.se, joregan@kth.se, edlund@speech.kth.se

Abstract

We present a shoal of voices in *Storspigg–TBI*, a legally cleared, professionally recorded Swedish speech corpus derived from talking-book production at the Swedish Agency for Accessible Media (MTM). The corpus contains 1 000 information messages read by 99 narrators under controlled studio conditions. The material has undergone full legal assessment and a three-sweep adoption process ensuring provenance, FAIR/FACT compliance, and reproducibility in collaboration with the national research infrastructure Språkbanken Tal. The paper describes the legal framework, data-selection and curation pipeline, as well as initial automatic transcription using Swedish Whisper and wav2vec 2.0 models. The resulting corpus provides a high-quality reference resource for speech science and technology, supporting research on inter-speaker variation, prosody, and evaluation under consistent acoustic and linguistic conditions.

Keywords: speech corpus, Swedish, controlled read speech, legality, reproducibility

1. Introduction

Speech science depends on reference corpora whose provenance, legal status, and recording conditions are fully documented. Yet most speech data today are large-scale, opportunistic collections with limited control over content or licensing.

Storspigg–TBI consists of short, uniform information messages read by nearly one hundred professional narrators. Message durations range from 34 to 75 seconds, primarily due to speaking rate and pausing. The material originates from the talking book production of the Swedish Agency for Accessible Media (MTM) and has been legally assessed and prepared for open research.

Controlled content and consistent recording conditions make the corpus uniquely suited for studying inter-speaker variation, prosody, and vocal characteristics. Such corpora also provide a stable and interpretable reference set for evaluating speech technology systems under known acoustic and communicative conditions, a type of material increasingly needed to ensure robust and meaningful performance assessment.

The main contributions of this paper are:

- A detailed account of the legal assessment that made these recordings available for research use.
- A reproducible pipeline for selecting and preparing the data.
- The adoption procedure used within the national research infrastructure Språkbanken Tal.
- The resulting corpus itself, including initial automatic transcription and descriptive statistics.

2. Background and related work

2.1. Multi-speaker speech corpora

Multi-speaker speech data have been collected in many forms and for many purposes over the past decades. For the purposes of this paper, we distinguish three broad categories that differ in how speech content, recording conditions, and speaker populations are controlled: (1) aggregated found data; (2) task-driven and/or limited domain multi-speaker corpora; and (3) controlled read-speech corpora. The grouping is not meant to be comprehensive or theoretically exhaustive, but rather a practical framework for situating the present work. Likewise, the examples cited below are illustrative rather than complete, with an inevitable bias toward widely used English and multilingual corpora, and, where available, Swedish material.

2.1.1. Crowdsourced, found and aggregated speech

Recent large-scale speech data initiatives have primarily relied on speech collected through open calls, public archives, or large-scale aggregation rather than through purpose-recorded, controlled corpus design. Examples include **Mozilla Common Voice** (Ardila et al., 2020), which collects prompted readings from volunteers worldwide, and the Finnish **Donate Speech** (Lahjoita puhetta) campaign (Lindén et al., 2022), a collaboration between Yle, the Finnish Language Bank, and Aalto University. The Finnish initiative gathered more than 4 000 hours of colloquial Finnish speech from over 25 000 speakers in only a few months, with legal and technical frameworks explicitly designed to permit both

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8446–8454
11–16 May 2026.

academic and commercial.

A related strand consists of large institutional or parliamentary recordings, such as **VoxPopuli** (Wang et al., 2021), covering more than twenty European languages, **NordParlTTS** (Li et al., 2026), covering Swedish and Finish, and the Swedish parliamentary archive, partially released as **RixVox** (Rekathati, 2023). These resources share the advantage of long temporal coverage and speaker diversity, but their research provenance or intended use, recording environments, and annotation depth vary greatly.

Finally, several projects combine and process speech from heterogeneous, often web-based sources into aggregated data sets. Examples include the **Spotify Podcasts Dataset** (Clifton et al., 2020), which collects transcribed spoken English from public podcasts; the **LibriSpeech** (Panayotov et al., 2015) and **LibriVox**[1], which are derived from volunteer audiobook readings; and the more recent **Emilia** pipeline (He et al., 2024), which systematically harvests and cleans large quantities of "found" audio such as podcasts, broadcasts, and online video. Despite their scale and diversity, such collections generally have limited control over recording setup, speaker metadata, or textual content. Their legal and copyright status is often as difficult to establish as their linguistic composition, making redistribution, benchmarking, or secondary use challenging. These resources thus stand in contrast to smaller, legally vetted corpora recorded under consistent and well-documented conditions.

2.1.2. Task-driven and/or limited-domain

Before the current combination of large-scale compute and machine-learning (ML) methods made it attractive to aggregate speech indiscriminately, large multi-speaker corpora were typically created for specific, well-defined research or technology-development purposes. These corpora were typically collected under controlled or well-documented conditions, but within a narrowly defined communicative setting such as telephone dialogues, meetings, read prompts, or scripted tasks. Examples include **SWITCHBOARD** (Godfrey and Holliman, 1992), **CALLHOME** (Canavan, Alexandra et al., 1997), the **AMI** (Carletta, 2007) and **ICSI** (Janin et al., 2003) meeting corpora.

Although these corpora are small by modern ML standards, their deliberate task structure, documentation, and annotation depth made them central to speech technology research and development for decades.

[1] https://librivox.org/

2.1.3. Controlled read-speech

The third and final category we discuss consists of corpora recorded under deliberately controlled conditions, where multiple speakers read the same or closely matched material. Such corpora were designed to minimise variability in linguistic content and recording setup, enabling systematic comparisons across speakers, dialects, and speaking styles. Examples include **TIMIT** (Garofolo et al., 1993), **SpeechDat** (van den Heuvel et al., 2001), and **Speecon** (Iskra et al., 2002), which together defined the model for large, phonetically balanced read-speech databases. Later corpora such as the **CMU Arctic** databases (Kominek and Black, 2003), **VCTK** (Yamagishi et al., 2019), and **LibriTTS** (Zen et al., 2019) followed similar principles, often targeting speech synthesis and voice conversion.

While these resources were primarily created for engineering purposes, they have since become essential reference material across speech science and related fields – including research on prosody, voice quality, speaker adaptation, and applied areas such as speech pathology and education. They illustrate a continued need for corpora in which quality, consistency, and clear documentation take precedence over scale or domain breadth.

Although not originally recorded for research, the corpus described in this paper largely fits this controlled tradition.

2.2. Actors in the process

2.2.1. Source and donor

The materials in **Storspigg–TBI** is donated by the Swedish Agency for Accessible Media (MTM), a Swedish government agency under the Ministry of Culture. One of its core missions is to adapt text, such as books, newspapers and educational materials, into accessible formats, for example easy-to-read text, Braille and speech. MTM also serves as a national knowledge centre for accessible media. In that role, MTM strives to participate in, and stay up to date with speech science and speech technology research and development, with an accessibility perspective. It has also long been active in the technical development of methods for talking book production, including the DAISY (Digital Accessible Information SYstem) format and early adoption of text-to-speech (TTS) production for university textbooks in the mid-2000s using fully in-house technology at a time when commercial Swedish TTS was not suitable for the task.

Further strengthening the motivation for MTM to actively contribute to Swedish speech science and speech technology research, the Swedish government announced in April 2023 its decision for Sweden to join the Open Data Charter (ODC; Directive

(EU) 2019/1024), an international framework promoting the accessibility and reuse of public-sector data[2]. The Swedish Agency for Digital Government (DIGG) is responsible for coordinating the implementation of the ODC nationally. In its guidelines, DIGG formally recommends that Swedish public-sector bodies make data of their own production available under open licences to the extent legally possible, and that they continuously strive for agreements and procedures that enable such sharing (Digg, 2025).

2.2.2. Maintainer and distributor

Storspigg–TBI is adopted, curated, maintained and distributed by the national research infrastructure Språkbanken Tal (Eng. roughly 'The Language Bank: Speech'). Språkbanken Tal is a national research infrastructure hosted by KTH Royal Institute of Technology. It forms the speech branch of the larger Språkbanken infrastructure and participates in CLARIN ERIC through its K-centre, CLARIN Speech.

One of Språkbanken Tal's core missions is to collect, adopt, curate, develop, maintain, and distribute resources for speech-oriented science, covering research into speech, speech technology, and fields in which speech plays a central role in investigation or application.

Språkbanken Tal manages a collection of related corpora under the collective name Storspigg (Eng. three-spined stickleback). These corpora share several defining characteristics:

- they contain a large number of items (e.g. recordings) from different sources (e.g. speakers) producing similar content;
- they are well-documented, consistently annotated, and recorded under controlled or known conditions;
- they are suitable for comparative studies and as reference material for evaluations and benchmarking.

All corpora adopted by Språkbanken Tal are distributed in stable, well-defined formats with persistent identifiers. In addition, they are transparently and systematically evaluated with respect to the FAIR (Findable, Accessible, Interoperable, Reusable) and FACT (Fairness, Accountability, Confidentiality, Transparency) principles, as well as data sensitivity and legal status. The evaluation is carried out in a methodical and standardised manner, and the resulting reports form part of each corpus's metadata. All Storspigg corpora are also fully adopted Språkbanken Tal corpora and meet the above standards. The corpus described in this paper is the first publicly released member of the Storspigg collection.

[2]https://eur-lex.europa.eu/legal-content/SV/TXT/?uri=CELEX:32019L1024

2.3. Legal issues

The legal landscape surrounding voice recordings in Sweden and the EU is complex. Since its inauguration in 2018, Språkbanken Tal has faced a series of obstacles to data sharing and is currently able to distribute speech data only under strict and limiting conditions.

For MTM, the situation is further complicated by a tension between its mission and EU's/DIGG's open-data recommendations on one hand, and the special considerations arising from MTM's unique legal standing on the other. MTM operates under an exception in the Swedish Copyright Act (Section 17), which allows the adaptation and distribution of accessible versions of copyrighted works without the rights holders' consent for people who cannot read printed text in its standard form. This exception requires MTM to handle the copyrighted material it adapts with particular care, and there is currently no fully sound legal mechanism for making its recorded literature available for research, even when generative-model training is explicitly excluded. Although the corpus presented here contains no copyrighted material, the clearance procedure has nevertheless been lengthy and demanding.

3. Method

3.1. Initial material selection

The initial stage of the work focused on identifying, isolating, and preparing a legally unproblematic subset of MTM's extensive holdings for corpus creation. Given the legal and ethical constraints, all materials containing copyrighted or potentially copyrighted content were excluded at the outset.

Since its foundation in 1980 as TPB (the Swedish Library of Talking Books and Braille), MTM has adapted tens of thousands of titles into formats such as Braille, easy-to-read text and speech, all of which are distributed through its digital library *Legimus*. Today, *Legimus* contains nearly 150 000 digital talking books, of which approximately 114 000 are in Swedish, 21 000 in English, and the remaining titles in other languages. Most titles are narrated by human voice talents, but since 2006, when MTM began using text-to-speech (TTS) for the adaptation of less complex Swedish and English university textbooks, an increasing number have been created with TTS. As of 2025, *Legimus* included around 17 000 titles produced with TTS, about 9 000 in Swedish and 8 000 in English. The

vast majority of the books contain an introductory message informing borrowers about production detail, copyright law and their responsibilities.

After attempts to find a legally sound solution for partial sharing of models or statistics based on the larger material proved unfeasible, attention shifted to these introductory information messages, which are owned by MTM and free of third-party rights. These messages constitute the basis for the selection used in the present corpus.

3.2. Legal procedure (MTM)

In 2019, MTM engaged an external law firm to determine whether any legal impediments existed to sharing the audio files for research purposes.

As a measure of extraordinary caution, all narrators who had been employed by subcontracted recording companies within the past ten years were informed (via their companies) that their narrations could be used for research purposes. The ten-year limit was chosen with reference to Article 14(5b) of the GDPR (EU, 2016), which allows exemptions from the obligation to provide information where doing so would be impossible or would entail a disproportionate effort, particularly in connection with scientific research. Extending the scope beyond ten years was considered disproportionate, as older records are often incomplete and narrators may no longer be traceable.

With these measures in place, the law firm saw no reason to withhold the files. Its recommendation was based on the following findings:

- **Information:** Narrators were informed that their recordings could be used for research.
- **File content:** The files consist mainly of factual data (page numbers, headings) without copyright protection. Excerpts from original works, however, must be removed if they occur in the extracted file.
- **Performers' rights:** Narrators' related rights under Section 45 of the Swedish Copyright Act are limited, as simple informational messages are generally not protected works.
- **Agreements:** MTM's framework agreement grants full economic rights to MTM, while narrators retain moral rights. Naming requirements are already met in the audio, and research publication is unlikely to be derogatory.
- **Risk assessment:** Even if related rights apply, the risk of infringement is low. Rights transfer is covered by the existing agreements.

MTM decision: Following the legal assessment, MTM decided to make the dataset *Information Files* available for research purposes under the Creative Commons Attribution–NonCommercial 4.0 International (CC BY-NC 4.0) licence in collaboration with Språkbanken Tal.

3.3. Data description

Generally speaking, the information files contain texts like this:

> *Information about Swedish copyright law and this talking book. This talking book has been produced for users of adapted media in accordance with section seventeen of Swedish copyright law. Illegal distribution or transmission will be prosecuted. The talking book is X pages long and has headings at Y level(s). It is produced by the Swedish Agency for Accessible Media, MTM, in YYYY. Narrated by NARRATOR at COMPANY. The e-text of the book is provided and can be accessed using text display software. End of information.*

In addition, the metadata in Table 1 is available for each file (some of the metadata is redacted for legal or legacy reasons).

3.3.1. Data selection procedure

As a starting point in 2018, MTM extracted the second audio file (the file assumed to contain the introductory information message) from each talking book adapted between approximately 2010 and August 2018 in *Legimus*. This yielded a total of 123 840 audio files, accompanied by the metadata shown in Table 1.

The target for the initial release of **Storspigg–TBI** was set to 1 000 Swedish audio files, as this was considered manageable from a curation standpoint while still large enough to be useful. Cascading filters were used to achieve the selection.

Filtering by language: All audio files in languages other than Swedish were removed, resulting in around 99 000 remaining files.

Filtering by date: Files produced with text-to-speech synthesis were removed (94 000 remaining). Files created before 2011 were then removed, as they most often contain the same pre-recorded information message read by the same narrator and without any book-specific information, such as the number of pages (19 000 remaining).

Filtering by media publisher: The list was filtered by media publisher: the company, library, authority or the like that adapted the written text material to a talking book. The vast majority of books were adapted by MTM (or by TPB, its former name),

Field	Description
Media number (redacted)	Book ID
Narrator	Name of narrator
Media type (redacted)	Talking book or Talking book with text
Title (redacted)	Book title
Author (redacted)	Author
Media publisher	Producer of adaptation
Media publication date	Date of adaptation
Target group	Adult or Juvenile
SAB code	The Swedish library classification scheme
Language	Language of the text
Course literature	University literature (Y/N)
Publication date	Date of publication of original book

Table 1: Metadata extracted for each audio file.
Fields not included in the shared corpus for legal or legacy reasons are marked "redacted".

but there are also books adapted by libraries or companies. Initial inspections suggested that the content of information file attached to adaptations by TPB/MTM was limited to information about the talking book, without any copyrighted text read aloud. Consequently, only these files were saved for the next round (15 000 remaining)

Filtering by message: The talking books belong to two main categories: *Talking book with text*, which generally contain the longer, above-quoted message read by the narrator of the book; and *Talking book*, which generally contain a shorter message; either a standard message of one speaker inserted in the beginning of the book or read by the narrator of the book. Only the *Talking book with text* type was kept (7 000 remaining).

Filtering by content: Files that may contain more than the information message (potentially copyrighted) were detected in two ways, resulting in 5 490 files remaining files:

1. ffmpeg was used to retrieve the length of the audio file, the file size, the sampling rate, and the bit rate. Audio files shorter than 15 seconds and longer than 75 seconds were considered to be not containing the target message and were skipped.

2. The audio files were then analysed using Whisper (Radford et al., 2022), to assure their content roughly matched the information message template presented above.

Chronological selection: The remaining files were sorted in chronological order and 1 000 files were selected approximately at equal distances.

3.4. Corpus adoption procedure

The corpus was incorporated into SBT through the infrastructure's structured *corpus adoption process*. This process ensures legal and ethical defensibility, technical immutability, and long-term scientific value.

3.4.1. Pre-adoption clearance

Legal and ethical conditions were evaluated separately from the technical workflow using **Gädda**, SBT's decision framework for legally and ethically defensible corpus adoption. Gädda introduces a structured assessment layer that allows legal and ethical clearance to be determined before a dataset enters the technical ingestion pipeline.

The framework operates by matching the assessed sensitivity level of a corpus with the technical and organisational measures (TOMs) implemented in the infrastructure. Sensitivity levels capture the degree of legal or ethical risk associated with a dataset, while TOMs describe the safeguards available for handling it, such as access controls, encryption, or restrictions on redistribution.

Adoption proceeds only if the corpus's sensitivity does not exceed the maximum level supported by the available TOMs. If the required safeguards are not available, ingestion cannot proceed. Gädda is inspired by the SEMLA model (Alexandersson et al., 22-12-14) and related approaches, but is adapted to Swedish legal and institutional conditions.

3.4.2. First sweep: registration

After clearance, the corpus was registered and assigned a descriptive name and a persistent identifier: (Talboksinformationskorpusen (TBI); Eng. the Talking-book information corpus) in the Storspigg collection. Each file was hashed with SHA-256, establishing its original content identity, and intrinsic properties such as size, format, and codec were recorded. The resulting identifiers and metadata were stored in authoritative SBT registries: the file

registry (intrinsic properties), the membership registry (corpus composition), and the provenance registry (derivations and curation steps).

3.4.3. Second sweep: non-lossy conversions; detailed format registration

Files were examined for compliance with the infrastructure's baseline requirements. When needed, they were minimally and non-destructively coerced into compliant versions of themselves – for example, by normalising audio containers or ensuring proper metadata encoding. The adopted originals were retained as the external origin, while the coerced versions became the internal canonical references.

All transformations and curation were recorded in the provenance registry, documenting each derivation and its relation to the original data. Alongside these authoritative records, the adoption process also produced non-authoritative but human-readable *sidecars* (XML), which replicate key registry information for inspection and portability. Together, these mechanisms guarantee transparency and enable sample-level reproducibility.

3.4.4. Third sweep: Initial transcription

The automatic transcription stage, referred to as *Sweep 3* in the adoption and curation workflow, applies ASR both as a processing and a validation tool. As with the preceding sweeps, all operations are logged in the provenance registry and are fully reproducible.

The audio files were automatically transcribed using two separate ASR systems. The main purpose of this step was to provide input for subsequent manual correction, producing a human-verified transcription with time alignments for each file. A secondary purpose was to obtain an early validation of the material through comparison between two independent ASR systems.

Each file was processed using both the **KB-Swedish Whisper** model and a Swedish **wav2vec 2.0** model. The resulting transcriptions were compared at the token level to produce insertion, deletion, and substitution counts and cumulative recognition tables (CRT). This comparison provides a lower-bound estimate of textual identity across the corpus and an indication of the typical segment length of identical material.

Importantly, no standard text normalisation was applied prior to inspection. Modern ASR systems often apply undocumented and inconsistent normalisation procedures that can obscure model-specific behaviour and make comparisons unreliable. Our approach instead aims to:

1. reveal idiosyncrasies and systematic characteristics of the models and tools themselves;
2. support the construction of a *vocabulary of near-equivalences* – cases where two strings differ in a Unicode sense but are functionally identical in the intended transcription target sense, identified through contextual examples; and
3. avoid reliance on opaque or opinionated preprocessing tools.

These decisions reflect the overall design principle of *human-in-the-loop curation*. Tools are used to accelerate processing, but humans evaluate every step to detect and characterise systematic errors. To support this process, lightweight web interfaces have been developed for rapid "same/not-same" judgements in context, a classifier tool that allows curators to assign canonical forms to equivalent pairs, and bespoke modules that propose likely matches and request explicit human confirmation.

Together, these components establish a reproducible, interpretable, and extensible framework for ASR-assisted corpus transcription and validation within Språkbanken Tal.

4. Results

As mentioned before, the main outcome of this work is the public release of the corpus **Storspigg–TBI** v1.0.0, the first member of the Storspigg collection. It contains 1 000 audio files of similar short texts, all read by professional voice talents, produced between 2010 and 2018. The recordings were curated, legally cleared, and adopted into Språkbanken Tal through the infrastructure's full three-sweep procedure described above. The corpus is distributed under the Creative Commons Attribution–NonCommercial 4.0 International (CC BY–NC 4.0) licence and is accessible through the Språkbanken Tal repository.

4.1. Corpus statistics

The final corpus comprises 1 000 Swedish audio files (12.7 hours total duration) from 99 professional narrators (51 female, reading 575 texts, and 48 male, reading 425 texts) spanning roughly one decade of production (2010–2018). The number of recordings per narrator ranges between 1 and 67, (average = 10.10, median = 5).

The original files are mono MP3-encoded audio, averaging 46 seconds in length, with a consistent bit rate of 48 kbps and a sampling rate of 22 050 Hz. They are accompanied by metadata including narrator, year, SAB (the Swedish library classification system), and target group.

The canonical adopted version of the corpus retains the same audio content but has been converted for consistency and long-term preserva-

tion. All files were converted to mono, 16-bit linear PCM WAV format at 22 050 Hz sampling rate and stored using lossless compression. File identifiers are neutral, registry-derived handles that carry no descriptive metadata, ensuring that file listings remain safe and that all sensitive or identifying information resides solely in the infrastructure's metadata registries. Pedigree and derivations between adopted, converted, and curated versions are recorded within a separate provenance system rather than in filenames.

No dynamic-range or amplitude normalisation was applied, preserving the original studio characteristics, but all files comply with the infrastructure's baseline requirements for archival stability and

4.2. Legal and procedural outcomes

The legal clearance procedure described in Section 3.2 resulted in a definitive assessment that the selected files are free from third-party rights and may be shared for research purposes. This represents the first instance of a Swedish public authority releasing structured voice recordings under a Creative Commons licence for sustained distribution through a research infrastructure. The process also tested and confirmed the applicability of the GDPR Article 14(5b) proportionality clause in the context of large-scale voice data, providing a precedent for future research collaborations with public bodies.

On the procedural side, the corpus constitutes a verified demonstration of the Språkbanken Tal *three-sweep adoption model*. All steps from initial legal evaluation through registration, non-lossy conversion, provenance documentation, and ASR validation were completed and logged within the infrastructure's registries, establishing a transparent, reproducible adoption trail.

4.3. Automatic transcription and validation

As part of the adoption process, each audio file was automatically transcribed using two independent ASR systems: **KB-Swedish Whisper** and a Swedish **wav2vec 2.0** model. The comparison between the two outputs served both as an early validation of audio quality and as a starting point for manual curation. A qualitative inspection of the outputs showed high overall correspondence between the systems, with most discrepancies involving punctuation, numerals, or inflectional variants. The analysis also revealed a small number of structural mismatches attributable to differing normalisation strategies and error handling between the ASR engines.

These results confirm that the recordings are clean, consistent, and suitable for detailed alignment and further study. The transcriptions, along with the comparison statistics and manual-correction interface, are included in the infrastructure's internal curation environment and will accompany the next public release once human verification is complete.

5. Discussion

The release of **Storspigg–TBI** v1.0.0 demonstrates that it is possible to publish controlled, professionally recorded speech data from a public-sector source under a clear legal framework. The collaboration between MTM and Språkbanken Tal shows that a balance between accessibility, data protection, and research openness can be reached through methodical, well-documented procedures rather than ad-hoc exceptions. In that sense, this work serves both as a corpus release and as a validation of the legal and technical mechanisms that enable such releases.

From the perspective of research infrastructure, the project provides a complete test case for Språkbanken Tal's adoption model. The layered approach – pre-adoption clearance, registration and non-lossy conversion, and ASR-assisted curation – proved effective for ensuring legal defensibility and scientific reproducibility. The process also showed that structured documentation of provenance and sensitivity assessment can transform the costly, opaque task of "making speech data shareable" into a transparent and repeatable procedure.

Scientifically, the resulting corpus contributes a new kind of resource to Swedish speech research: a multi-speaker collection of short, homogeneous recordings produced under professional studio conditions and free from third-party rights. It complements existing large-scale but uncontrolled data sets with a smaller, legally robust, and experimentally interpretable reference corpus. The material is immediately applicable to studies of inter-speaker variation, prosody, and voice characteristics, and to the development and evaluation of speech-technology systems under consistent conditions.

Finally, the case highlights the need for sustained coordination between public authorities, legal experts, and research infrastructures. The procedural effort required to make even legally uncomplicated recordings available for research illustrates that legal and ethical reproducibility must be considered integral to corpus creation. Future Storspigg corpora will benefit from this experience, reducing the time and resources required for each subsequent adoption and strengthening the long-term sustainability of Swedish speech resources.

6. Future work

Future work will extend both the material and the processing of the corpus. Planned activities include:

- collecting additional information messages produced between 2018 and the present day;
- creating a parallel corpus of English information texts recorded under comparable conditions;
- extending the current corpus with exact orthographic and phonetic transcriptions and time alignments; and
- developing standardised test and evaluation data sets based on the curated material.

7. Conclusions

The Storspigg–TBI corpus demonstrates that large-scale, high-quality spoken resources can be created and shared within a rigorous legal, ethical, and infrastructural framework. Unlike the uncontrolled or weakly documented collections often used for machine-learning training, this corpus combines professional recording standards with complete provenance, transparent licensing, and reproducible curation. Through its adoption into Språkbanken Tal, it establishes a repeatable model for transforming restricted institutional holdings into open, research-ready data.

Beyond its immediate scientific use, the work validates a practical path toward sustainable speech resources: controlled content, traceable procedures, and lawful openness. By showing that reproducibility and legality need not come at the expense of usability, it argues for a new generation of reference corpora that complement large, opportunistic data sets with smaller, well-understood, and responsibly shared materials.

Acknowledgements

The results of this work will be made more widely accessible through the Swedish Research Council funded national infrastructure Språkbanken Tal (2023-00161_VR).

8. References

Jan Alexandersson, Jochen Britz, Valentin Seimetz, and Daniel Tabellion. 22-12-14. SEMLA: An on-premises trusted research environment for AI-based R&D with sensitive personal information. White Paper SEMLA v1.0.0, DFKI.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Canavan, Alexandra, Graff, David, and Zipperlen, George. 1997. CALLHOME American English Speech. Artwork Size: 1830160 KB Pages: 1830160 KB.

Jean Carletta. 2007. Unleashing the killer corpus: Experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation*, 41(2):181–190.

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. 100,000 Podcasts: A Spoken English Document Corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Digg. 2025. Rekommendation om öppna licenser och immaterialrätt | Digg.

EU. 2016. General Data Protection Regulation 2016/679.

John S. Garofolo, Lori F. Lamel, William M. Fisher, David S. Pallett, Nancy L. Dahlgren, Victor Zue, and Jonathan G. Fiscus. 1993. TIMIT Acoustic-Phonetic Continuous Speech Corpus. ZSCC: NoCitationData[s1].

JJ Godfrey and EC Holliman. 1992. SWITCHBOARD: Telephone speech corpus for research and development. *Acoustics, Speech, and Signal Processing*.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, and Peiyang Shi. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *SLT 2024*, pages 885–890, Macao, China. IEEE.

Dorota Iskra, Beate Grosskopf, Krzysztof Marasek, Henk van den Heuvel, Frank Diehl, and Andreas Kiessling. 2002. SPEECON – speech databases for consumer devices: database specification and validation. In *Proceedings of the Third International Conference on Language Resources*

and Evaluation (LREC'02), Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Adam Janin, Don Baron, Jane Edwards, Daniel P. W. Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI Meeting Corpus. In *Procs. of ICASSP'03*, pages 1520–6149, Hong Kong, China. IEEE.

John Kominek and Alan W. Black. 2003. CMU Arctic databases for speech synthesis. Technical report, Carnegie Mellon University, Pittsburgh, USA.

Zirui Li, Jens Edlund, Yicheng Gu, Nhan Phan, Lauri Juvela, and Mikko Kurimo. 2026. Nord-Parl-TTS: Finnish and Swedish TTS dataset from parliament speech.

Krister Lindén, Tommi Jauhiainen, Mietta Lennes, Mikko Kurimo, Aleksi Rossi, Tommi Kurki, and Olli Pitkänen. 2022. Donate Speech. In Darja Fišer and Andreas Witt, editors, *CLARIN: the Infrastructure for Language Resources*, pages 481–510. De Gruyter.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, South Brisbane, Queensland, Australia. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Faton Rekathati. 2023. The kblab blog: Rixvox: A swedish speech corpus with 5500 hours of speech from parliamentary debates.

Henk van den Heuvel, Louis Boves, Asuncion Moreno, Maurizio Omologo, Gaël Richard, and Eric Sanders. 2001. Annotation in the SpeechDat projects. *International Journal of Speech Technology*, 4(2):127–143.

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: a large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Procs. of ACL-IJCNLP 2021*. Association for Computational Linguistics.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92).

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. ArXiv:1904.02882 [cs].